

Human genomes, public and private

The burgeoning commercial sector that is based on genome information poses a challenge to the norms of scientific publication. But it remains to be established that the conditions of access to published sequence data need to change.

The human genome sequence contains the genetic code that sits at the core of every one of the ten trillion cells in each human being. It profoundly influences our bodies, our behaviour and our minds; it will help the study of non-genetic influences on human development; it will unlock new insights into our origins and history as a species; and it points to new ways of combating disease. The people of many countries have invested in the Human Genome Project's determination of the sequence, and it is hard to see how that investment could have received better returns. Having released their data daily from the outset with unrestricted access, the publicly funded consortium has assembled about 92% of the sequence. *Nature* is delighted this week to publish the project's analysis, and related results, freely available to all without restriction at <http://www.nature.com> and on pages 814–958.

In so doing, *Nature* has followed a traditional model in the publishing of extensive scientific data. As indicated in our "Guide to Authors", we require the results of genome sequence analyses, as with protein structure coordinates, to be immediately available from an appropriate database without restriction. This supports an unwritten contract with our readers that what they see described is what they can use, without obstacles, whether they work in the commercial or academic sector (an increasingly blurred distinction). It supports a broader principle by which scientific results are available for searching and use with software tools. And it supports a principle enunciated by the United Nations that the human genome in particular is, in a symbolic sense, humanity's common heritage.

It is worth noting that *Nature* sees a distinction between access to essential data embodied in a paper, and access to the materials and techniques used. Making freely available all the materials used for a piece of research is sometimes impractical. All we can do is to sustain a policy of access to materials as far as possible, and to cooperate in attempts to set up new standards of access as technologies develop.

Money makers

What, then, of companies that make a living from data discovered through their own research? There is big money at stake. The bioinformatics market alone will soon exceed \$1 billion per year. Some pharmaceutical companies are valued at hundreds of billions of dollars, much of which depends on their ability to exploit genomic information in developing blockbuster drugs. According to stock-market commentators, there are more than 100 drug-platform biotech companies chasing some \$3 billion of research and development money. Genomics-based drug companies such as Human Genome Sciences, Millennium and CuraGen are forming partnerships with big drug companies to help them grow. Companies such as Affymetrix, Celera, Gene Logic and Incyte are spending billions on software and hardware. Information-technology giants such as Motorola and IBM are increasingly bringing their expertise to bear in the supply and analysis of microarrays and other biochips.

How do information generators and curating companies make their living? Celera, a company in the vanguard of those seeking to use the human genome sequence to make money for private investors, depends on much more than the sequence to pay its way. Its sources of revenue include subscriptions, software, third-party licence revenue, consulting and diagnostics. Its total revenues are estimated at \$130 million, and its capitalization stands at over \$1 billion. Some analysts are concerned about Celera's future profitability in the face of the availability of free data from public projects. Others are confident that it will be able to add value to the basic information in a way that drug companies will pay substantial sums for. The company is gearing up to become a proteomics powerhouse and, as its president Craig Venter has said, there are too many challenges around for it to make sense for private and public scientists to duplicate one another's work.

Data access

The private sector is essential to the improvement of human health, and its investors need financial returns. Yet many of the commercial companies, whatever their motives and products, do outstanding basic science along the way. The amount of valuable scientific information residing in research and development centres in the private sector is huge — the global R&D budget of pharmaceutical companies is estimated at \$28 billion per year.

It is in the interests of all researchers that as much of that information as possible finds its way into peer-reviewed publications. Suppose a journal were handed a wonderful and unique piece of science, but that, in the interests of the originators, some of the information had to be kept behind a wall of subscriptions and licensing arrangements. In that case the conflict of interest would be acute: on the one hand, the world is better served if the information is available, even with conditions, than if it is not available at all; on the other hand, such a situation not only breaches the traditional norms of scientific publication, but also hampers the ability of science to progress rapidly by the unrestricted reanalysis of published data.

Since we established our policy on access to genome data in January 1996, *Nature* has been able to hold the traditional line. The burden of providing proof that the line should be abandoned lies with the companies — either through rational debate or possibly by the sheer scientific significance of their output in the absence of a publicly funded equivalent. With a publicly funded project delivering data, *Nature* believes that the human genome sequence is not the place for the traditional rules to be broken. We are willing to cooperate with competing journals to maintain the model of open access as far as possible, as we have in the past. We are also willing to engage in consultation where new types of privately generated data, lacking their equivalent in the public domain or in standard public databases, require new thinking. But, in the meantime, the more that both private and public activities can stimulate each other with freely available science, the better it will be for everyone. ■





Mapped out
The human genome is unveiled in all its glory
p747



Crop control
Experts call for tighter regulations on transgenics
p749



Ray of hope
Cut-price anti-AIDS drugs offered to developing world
p751



Food for thought
Public consortium anticipates early finish of rice genome
p752

Publication of human genomes sparks fresh sequence debate



Declan Butler

This week's publication of the human genome sequence by both Celera Genomics of Rockville, Maryland, and the publicly funded international

Human Genome Project (HGP) has reignited the debate over the relative merits of the two teams' different strategies.

The two groups published their work simultaneously, as promised last summer, and held a cordial joint press conference in Washington on Monday to advertise the fact. At five more press conferences around the world, participants in the public project celebrated their achievement, which is published in *Nature* (see pages 860–921).

But in the run-up to these meetings, leading members of both teams had been working hard in an attempt to ensure that history — or at least the media — would judge them to have made the more important contribution.

Celera, which embarked on its sequencing effort only three years ago, needs to convince customers who will pay for access to its database that what they are getting is superior to the freely available public data. Members of



Cordial: Celera's Craig Venter (left) and the HGP's Francis Collins at Monday's press conference.

the HGP, having fought off what they regard as an effort by Celera to undermine their project, are now arguing that Celera's assembly, published in *Science* (291, 1304–1351; 2001), could not have been completed without drawing heavily on their own work.

The public project adopted a 'clone-by-clone' approach. In this, the entire genome is chopped into chunks up to several hundred thousand base pairs long, and inserted into synthetic chromosomes known as bacterial artificial chromosomes (BACs).

The key to the HGP's strategy is the subsequent 'mapping' step in which the BACs are each positioned on the genome's chromosomes by looking for distinctive marker sequences, called sequence tagged sites (STSs), whose location has already been pinpointed. In this way, the BACs provide a high-resolution map of the entire genome.

Clones of the BACs are then shattered into tiny fragments in a process known as shotgunning. Each fragment is sequenced and computer algorithms that recognize matching sequence information from overlapping fragments are used to reconstruct the complete sequence inserted into each BAC.

But in 1997, Gene Myers, now vice-president of informatics research with Celera, and James Weber of the Marshfield Medical Research Foundation in Wisconsin argued that the mapping step was unnecessary. They said that algorithms used to reassemble shotgunned DNA fragments could be applied to cloned random fragments taken from the genome as a whole (*Genome Res.* 7, 401–409; ▶

And now for the proteome...

Alison Abbott

In an announcement timed to coincide with the publication of the human genome sequence, a group of top-level proteomics researchers has launched a global Human Proteome Organisation (HUPO).

HUPO's founders see it as a post-genomic analogue of the Human Genome Organisation (HUGO). Its mission will be to increase awareness of, and support for, large-scale protein analysis, in scientific, political and financial circles.

HUGO was created in 1988 by publicly funded researchers who wanted to coordinate global efforts to sequence the genome. Now that the draft human genome sequence has been published, researchers are turning their attention to identifying the functions and expression patterns of the proteins encoded

by the genes. It is generally believed that patterns of protein production — the proteome — will correlate with disease states, which may lead to new treatments.

"Proteins are central to our understanding of cellular function and disease processes, and without a concerted effort in proteomics, the fruits of genomics will go unrealized," says Ian Humphrey-Smith of the University of Utrecht in the Netherlands, one of HUPO's founder members.

So far, the embryonic HUPO has created a Global Advisory Council to foster international cooperation, and two regional task-forces in Europe and Japan. An inaugural meeting will take place in the spring to define detailed objectives and to elect a president.

1997). In this 'whole genome shotgun' strategy, fragments are first assembled by algorithms into larger scaffolds. The correct position of these scaffolds on the genome is then worked out using STSs.

Although the whole genome shotgun strategy had been successfully applied to small, simple genomes — such as those of viruses and bacteria — critics argued that it would not work for the human genome, which contains millions of repetitive DNA sequences. The critics expected these repeats to confound the algorithms, making a complete genome assembly impossible.

Nonetheless, Celera was founded with the mission of solving the human genome sequence in short order using a whole genome shotgun approach. The public project's criticisms of the Celera paper have focused on the company's alleged failure to meet this goal.

Data disputes

Eric Lander, director of the genome centre at the Whitehead Institute for Biomedical Research at the Massachusetts Institute of Technology, asserts that the Celera project would have been unable to find locations for much of its sequence without reference to the public project's genome map.

Celera's paper actually describes two genome assemblies, one put together using the whole genome shotgun approach and a second, 'compartmentalized' assembly. Both were done using Celera sequencing data in which each base of DNA had been sequenced, on average, 5.1 times (5.1X coverage), plus, the paper says, a further 2.9X coverage taken from the HGP's publicly available data.

But members of the public project say that this description is misleading. They argue that the 2.9X coverage is not a random selection of the HGP data. Instead, they say that the data were carefully chosen to cover the entire genome, giving few gaps and retaining maximal mapping information. As a result, they argue, Celera actually obtained the equivalent of 7.5X coverage from the HGP's data.

Lander also notes that Celera's whole genome shotgun sequence contains almost 119,000 scaffolds, rather than the 5,000 that the company had originally predicted. It is impossible to position this many scaffolds on the genome using STS markers, he argues. The majority of Celera's scaffolds are very small, claims Lander, and represent a "tossed genome salad".

Celera's compartmentalized assembly put the genome together region by region, making some use of the HGP's mapping information. In theory, this combination of both projects' sequencing data should produce a better sequence than the HGP data alone. But leaders of the public project argue that there is actually very little in it. "Remarkably, this product is very similar to ours," says John Sulston, former director of Britain's Sanger Centre near Cambridge.



Shooting from the hip: Celera claims its assembly methods have produced a more accurate genome.

"For three years the public project was told that we were inefficient, slow and pointless for proceeding in a careful fashion, and that a whole genome shotgun obviated the need for all these 'wasteful' steps," Lander says. "At the end of the day, it has transpired that we have been the ones who have guaranteed that there is a human genome sequence. We have saved the day," Sulston adds: "They [Celera] failed by their own standards."

Myers dismisses the criticisms from the HGP members as "pure speculation". The whole genome shotgun approach "worked extremely well", he says. "We got tremendous continuity and order and orientation across the genome."

Myers agrees that the method produced a large number of scaffolds in total, but says that more than 90% of the genome sequence is contained in less than 3,000 scaffolds. And he disputes the argument that the contribution from the public project was worth more than 2.9X coverage. "If we had had 3X more Celera sequence data, we could have done it completely on our own," he says.

Originally, Celera had planned to generate its own sequence with 10X coverage. Craig Venter, Celera's president, says that the company decided to make use of the public data, once it became clear that it would be available in time, "instead of spending six months and [another] \$60 million".

Strategic sequences

Myers points out that the publicly funded mouse genome project is now planning to use whole genome shotgun techniques, in addition to generating a BAC map. Thanks to Celera's efforts, says Myers, strategies for sequencing large genomes have been revised. "I'm proud of being a part of that," he says.

Celera has also responded with a critique of the HGP's sequence. At a press briefing last week, members of the Celera team outlined comparisons of the two papers which, they argued, showed that the company's sequence is more accurate than that produced by the HGP. Myers described analyses of 'mate pairs', which correspond to sequences from either end of an individual cloned DNA segment.

The length of these segments is known. So if mate pairs subsequently end up the wrong distance apart within the final sequence, it suggests a problem with the assembly. Myers pointed out that, for certain chromosomes, the HGP's sequence contains many more of these 'break points' than does Celera's. For chromosome 1, for example, he said, there are "about 35 times more break points in the public assembly".

But Richard Durbin, deputy director of the Sanger Centre, responds that mate-pair analysis is an integral part of Celera's assembly strategy, so one would expect it to make the Celera sequence look good. Nevertheless, Durbin acknowledges that, at present, the detailed ordering in parts of the Celera's sequence may be superior — especially in the case of chromosomes such as chromosome 1, which is still largely in draft form.

The dispute over the success or otherwise of the respective approaches is set to resonate for months, or even years. It will be fanned by the recognition that Nobel prizes may be at stake, and its resolution hampered by the fact that the intricacies of genome assembly are fully understood by only a small community of experts — most of whom have a foot firmly in either the public or the Celera camp.

But Ari Patrinos, head of biological and environmental research at the US Department of Energy, who over the past two years has been a tireless peacemaker between the two rival projects, hopes that publication of the two sequences will lead to a more considered and constructive analysis of the different methodologies by competent experts.

To aid this process, Patrinos intends to organize a joint Celera-HGP workshop, probably to be held in Washington on 3 April. The plan is for the meeting to be chaired jointly by Myers and David Haussler, an expert in computational biology at the University of California at Santa Cruz. Its goal, says Patrinos, is to establish "what can we learn from the experiences of both sides, and what in the future should be the optimum approach to sequencing mammalian genomes". ■

Additional reporting by Peter Aldhous in London and Colin Macilwain in Washington.

Call for tighter controls on transgenic foods

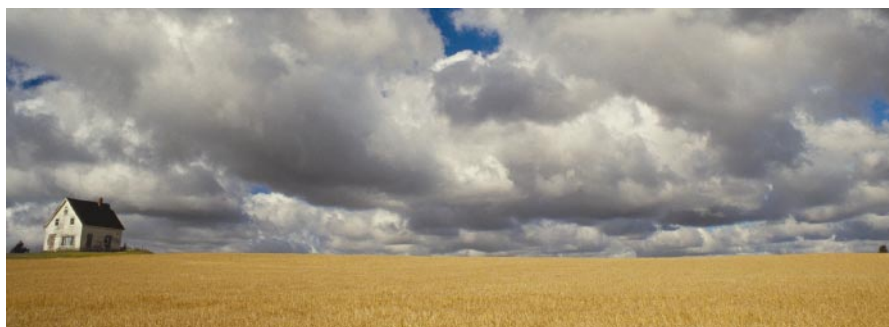
David Spurgeon, Montreal

A senior panel of scientific experts has surprised Canadian regulators — and many Canadians — by calling for far tighter regulation of genetically modified (GM) foods.

The recommendations made by the panel from the Royal Society of Canada (RSC) include a ban on growing GM fish in coastal netpens. Most importantly, the panel rejects the doctrine of 'substantial equivalence', by which regulators treat the approval of GM crops as though they were much the same as conventionally grown crops.

The panel says that approval of new transgenic organisms for release into the environment or as food should be based on "rigorous scientific assessment of their potential for causing harm". Such tests "should replace the current regulatory reliance on 'substantial equivalence' as a decision threshold", it declares.

The RSC panel, which prepared its report for the Canadian government's scientific and



Equal rights? GM crops should not be treated as 'substantially equivalent' to other plants, says panel.

regulatory agencies, says that the testing should be done in "open consultation" with the scientific community and that the results should be monitored, in public, by a panel of "experts from all sectors".

New technologies should not be presumed safe unless there is a reliable scientific basis for considering them to be so, the report adds. And "the primary burden of

proof [should] be upon those who would deploy food biotechnology products to carry out the full range of tests necessary to demonstrate reliably that they do not pose unacceptable risks".

The best scientific methods should be used to reduce uncertainties about risks to human health, the ecosystem and biodiversity, and "approval of products with these potentially serious risks should await the reduction of scientific uncertainty to minimum levels", the report adds.

The panel also says that potential environmental risks posed by GM fish should be assessed on a population-by-population basis. It adds that comprehensive research on the interactions between wild and farmed fish is needed before the risks posed by GM fish can be assessed.

The report contends that regulatory agencies should be more transparent about the science on which their decisions are based, and should take care to maintain a neutral stance in their public statements about the risks and benefits of biotechnology.

It also calls on the Canadian Biotechnology Advisory Commission to "review problems related to the increasing domination of the public research agenda by private, commercial interests".

The recommendations have been widely interpreted in Canada as a reprimand for the government's previous readiness to approve GM crops. But Conrad Brunk, academic dean at Conrad Grebel College at the University of Waterloo, Ontario, and a co-chair of the panel, claims that they are not very different from recent statements made in the United States and the European Union on the safety of GM foods.

"We were asked to forecast the directions of the technologies as they become more sophisticated and complex, and to make recommendations about the scientific capacity that would be needed to regulate them," he says. "Unfortunately, our recommendations are being interpreted by some people in the Canadian regulatory agencies as severe criticism of what they're doing now."

http://www.rsc.ca/foodbiotechnology/indexEN.html

Physicists worried by grant reforms

David Adam, London

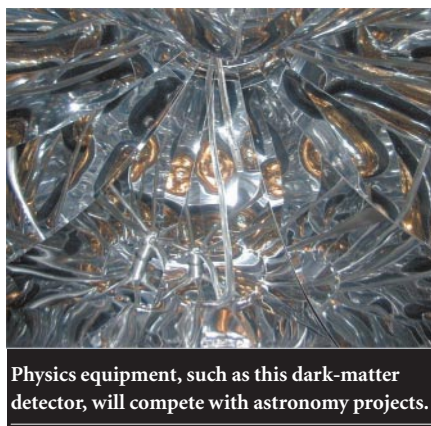
Britain's Particle Physics and Astronomy Research Council (PPARC) is to require high-energy physicists and astronomers to compete head-to-head for project funding for the first time.

The PPARC was formed in 1994 as the main UK research agency for two disciplines which each rely heavily on the long-term planning of expensive facilities. But until now, its advisory and peer-review structure has been firmly divided between separate arms for astronomy and high-energy physics.

Some high-energy physicists have expressed concerns about the new structure, which is expected to be approved by the PPARC's governing council this week. The physicists fear that their projects, which often involve years of experiment design and construction, may lose out in direct competition with equally expensive but more rapidly productive astronomy projects.

Under the new structure, a single panel will review and approve small project grants in both astronomy and high-energy physics, helped by appropriate specialists. Funds for larger projects, costing up to £3 million (US\$4.4 million), will be controlled by the council's main scientific committee.

John Garvey, a high-energy physicist at the University of Birmingham and a member of the PPARC panel convened to review its structure, says that his peers' earlier concerns have been addressed by the latest plan. "We now have something on the table



Physics equipment, such as this dark-matter detector, will compete with astronomy projects.

that everyone broadly agrees to," he says. "But at the beginning I do think there was a lack of understanding that particle physicists are experimentalists and astronomers are observers, and that the problems we face pursuing our science are different," he adds. The high-energy physicists advising the panel will now serve for three years, bringing more continuity, Garvey says.

"This is a change, and change gets people worried," admits PPARC chief executive Ian Halliday. "But I think the fears of some in the particle-physics community are unjustified. We don't have any hidden agendas." He acknowledges that the current system is effective, but says that the new structure will "bring greater focus, consistency and transparency to decision-making."

Reforms woo scientists from mainland China

David Cyranoski, Tokyo

Taiwan is taking several tentative steps to open doors to researchers from the People's Republic of China and overcome half a century of scientific estrangement between the two.

The latest move to improve scientific relations came last week when Taiwan increased the maximum time that visiting researchers from the mainland can stay on the island from two years to three.

Other changes expected in the near future include the introduction of multiple-entry visas for visitors and new provision for the direct transfer of money to banks on the mainland, says Kuan-Hsiu Hsiao, an official at Taiwan's National Science Council.

Researchers say that multiple-entry visas are essential to encourage scientific exchange. Currently, visitors from the mainland must go through a laborious process, taking several months, to obtain a limited travel permit for Taiwan. But the permit does not allow travel outside Taiwan — to visit international conferences or laboratories abroad — greatly restricting the mobility of those mainland researchers who do spend time on the island.

There are fewer restrictions for researchers wishing to move in the other direction, although few Taiwanese go to the mainland to do research. But there is an increasing desire to develop collaborations as many Taiwanese note great improvements in scientific practices on the mainland. "When I visited five years ago, they just wanted to show build-



ings and machines. Now they want to discuss scientific methods and results," says one researcher with the Academia Sinica, Taiwan's premier research organization.

Some of the restrictions are relatively minor: rules against depositing money into mainland banks, for example, just mean that visiting researchers have to find a Taiwanese friend to make the transfer.

Nonetheless, the latest changes, in combination with others made late last year, will produce a more agreeable atmosphere for scientific exchange, researchers say. Last December, Taiwan implemented a reform to allow mainland-born spouses of Taiwanese nationals to work at universities and public research institutes. Previously, they needed two years of residency in Taiwan and eight years of residency outside mainland China to qualify.

And for the first time, mainland researchers working in Taiwan for more than four months will become eligible for Taiwan's national insurance, with the Academia Sinica required to ensure insurance coverage for mainlanders. This change is a response to an incident in which a visiting postdoctoral student amassed huge medical bills after being hit by a drunk driver and spending three months in hospital in a coma.

But a bottleneck remains for those seeking permanent positions. "Undefined restrictions leave no pathway for mainlanders to find a job in Taiwan except in very special cases," says an Academia Sinica researcher. Taiwan, for example, still does not recognize degrees from mainland universities when considering people for institute and university research positions.

The mainland also imposes some obstacles. Travel permits for researchers are often granted for only six months or one year — barely enough time to get started on a research project.

According to a mainlander at one Academia Sinica institute, there are many reasons researchers from mainland China might want to go to Taiwan, including higher salaries, better facilities and a more international research environment. Taiwanese institutes and universities also face a dearth of qualified personnel and have a hard time attracting people from abroad because of language and cultural barriers. "If we make it easy, gradually [mainland researchers] will come," says Kuan-Hsiu Hsiao. ■

France and Spain join forces over synchrotron project

Xavier Bosch, Barcelona

Spain has agreed to help build a new synchrotron which is being planned outside Paris, in exchange for French participation in an as-yet unspecified major Spanish research facility.



Mutual benefit: Schwartzberg (left) and Birulés sign the science agreement.

The Spanish and French science ministers, Anna Birulés and Roger-Gérard Schwartzberg, reached an agreement in Madrid on 5 February to cooperate on the construction and operation of large research facilities. While Spain helps to build the Soleil synchrotron, it will consider various projects for French participation.

Neither country has disclosed the size of the Spanish contribution to the synchrotron, but a French newspaper has estimated it at FF170 million (US\$24 million) towards a total cost of about \$220 million. Schwartzberg told a press conference that Spanish researchers would join the scientific advisory board of the Soleil synchrotron.

Candidates for the Spanish facility, meanwhile, include a research vessel, a 10-metre optical telescope in the Canary Islands, and another synchrotron.

Spain appointed a committee last month

to study the feasibility of building its own synchrotron, and is expected to decide whether or not to go ahead with the project by the summer.

Two years ago, a design study for a synchrotron facility to be housed at the Autonomous University of Barcelona was positively evaluated and its construction recommended by the European Science Foundation (see *Nature* 405, 604; 2000).

Jérémie Boussand, the science attaché at the French Embassy in Madrid, says that if Spain decides to proceed with the synchrotron, then that will be the project that France will support.

Synchrotron light sources, which enable researchers to closely examine three-dimensional structures, have emerged in recent years as an important tool in materials science, chemistry, condensed-matter physics and structural biology. ■

♦ <http://www.ils.ifaef.es>

Indian company offers cheap anti-AIDS drugs



Some hope: but Western pharmaceutical companies are expected to block the cut-price drugs.

David Dickson

An Indian drug company has offered to supply a triple-drug anti-AIDS 'cocktail' to sufferers in developing countries at less than one-twentieth of the standard cost in the West.

Cipla of Mumbai says it has developed its own methods for producing the three drugs — stavudine, lamivudine and nevirapine — thereby avoiding the process patents owned by the major manufacturers in India.

The company says it will supply a combination of the drugs to the French charity Médecins sans Frontières (Doctors without Borders) for \$350 a year per patient, on condition that the organization provides the treatment free to patients.

Cipla has also offered the combination directly to wholesalers at \$1,200 and to governments at \$600 a year. "We're not making money, but we are not going to lose money either, since with the average of the three prices we should break even," Yusuf Hamied, the chairman of Cipla, told *The Times of India* last week.

Hamied said the move was inspired by the suffering caused by the recent earthquake in Gujarat, and predicted that AIDS was going to be "a bigger holocaust in India than the earthquake".

The action is being seen as a bid to force the major producers of anti-AIDS drugs to reduce their prices in developing countries to an affordable level. But it could fall foul of allegations of patent infringement. At present, the standard annual cost of the triple therapy in the United States and Europe is between \$10,000 and \$15,000.

Last year, the five leading drugs manufacturers agreed through the World Health Organization to negotiate with individual countries on a preferential pricing strategy. Deals have already been reached with Uganda, Senegal and Rwanda. Although no prices have been released, a figure of \$1,000 for annual treatment has been reported for Senegal.

Bristol-Myers Squibb, GlaxoSmithKline (created last December from a merger between Glaxo Wellcome and SmithKline Beecham) and Boehringer-Ingelheim are the main manufacturers and patent holders of

the three drugs. Although the companies have made no public comment on Cipla's move, they are expected to vigorously oppose what they will regard as an attempt to undermine their drug patents.

Last December, for example, Cipla stopped importing its own low-cost combination of the anti-AIDS compounds lamivudine and zidovudine into Ghana after being warned by Glaxo Wellcome that it was infringing the company's patent rights.

Cipla's offer is already being used by the British aid charity Oxfam to highlight what it describes as an "undeclared drugs war" by the pharmaceutical industry and governments of rich nations against the world's poor.

On Monday, the charity announced the launch of a campaign in which it is demanding that companies contribute a percentage of the profits from any drugs which earn more than US\$1 billion per year to a research fund for diseases that afflict poor people. ■

Fears grow over melting permafrost

Quirin Schiermeier

Melting permafrost in Arctic regions is likely to accelerate global warming, as well as disrupting the lifestyle of indigenous people, researchers warned last week's meeting of the United Nations Environment Programme (UNEP) in Nairobi.

A seventh of the Earth's carbon is stored in frozen Arctic soil, the scientists say, and huge amounts of greenhouse gases will be released into the atmosphere if rising temperatures cause the permafrost to melt and its organic material to be broken down by bacteria.

"There is now evidence that the permafrost in some areas is starting to give back its carbon," Svein Tveitdal, director of UNEP's Global Resource Information Database in Arendal, Norway, told a meeting of UNEP's governing council.

Tveitdal also pointed to regional problems likely to be caused in Siberia, Scandinavia, northern Canada and Alaska by melting permafrost, which can trigger subsidence and damage buildings.

UNEP scientists fear that rising temperatures and melting of the permafrost may have an impact on Arctic wildlife such as reindeer, and on the traditional lifestyle of indigenous people in northern regions.

"In Canada's north, we are seeing dramatic changes that affect permafrost and sea ice, the latter of which has major implications for species on which the traditional Inuit life depends, such as polar bears and seals," said David Anderson, Canada's environment minister and

president of UNEP's governing council.

Delegates at the meeting criticized the lack of a unanimous political response to global climate change. "We are taking action domestically, but we need awareness and movement on the international front as well," Anderson said.

The UNEP governing council is the main international forum for governments to address environmental policy issues. The meeting was attended by environment ministers from over 100 nations, along with scientists and representatives of non-governmental organizations and corporations.

Klaus Töpfer, executive director of UNEP, told the meeting that the time for talk had ended, and the time had come for delivering environmental action. "We do not need new priorities or new visions," he told the Nairobi meeting. "What we need to do now is to implement." ■



Melting moment: as the permafrost thaws it releases greenhouse gases into the atmosphere.

Rice genome consortium will finish ahead of schedule

Tokyo The International Rice Genome Sequencing Project has announced that it will finish sequencing of the *Oryza japonica* rice genome by the end of this year.

The publicly funded, 10-country consortium, led by a Japanese research team, had previously targeted 2004 for completing its reading of the estimated 430 million bases of DNA that make up the rice genome.

But last month a private company announced that it had completed the sequence (see *Nature* **409**, 551; 2001).



Fast food: the publicly funded group sequencing the rice genome is speeding up its effort.

Leaders of the consortium say they must finish the project to ensure that the data are accurate and freely available.

Transgenic plant patents get the go-ahead

Munich The European Patent Office's appeals board has confirmed the validity of two controversial patents on transgenic plants.

The board overruled objections filed by environmentalist groups against patents on a herbicide-resistant plant developed by AgrEvo and Monsanto's 'Flavr Savr' tomato.

The protesters had argued that the European Patent Convention prevented plant varieties from being patented.

But last year the patent office lifted a four-year moratorium on applications for patents on plants (see *Nature* **403**, 3; 2000). After last week's decision, environmentalists criticized the office for allegedly protracting its decision until its patenting rules were more favourable to the biotechnology industry.

Anthropology book's claims win inquiry

San Diego The American Anthropological Association (AAA) will conduct a full-scale inquiry into a book's allegations of scientific

misconduct and ethical lapses during years of research on the Yanomami tribe in Venezuela.

A five-strong task-force chaired by University of Arizona anthropologist Jane Hill will examine research concerns raised by Patrick Tierney in his recently published book *Darkness in El Dorado: How Scientists and Journalists Devastated the Amazon* (see *Nature* **408**, 755; 2000).

The AAA has also asked its ethics committee to develop draft guidelines for anthropologists in the light of issues Tierney raised concerning interactions between anthropologists and the people they were studying. The guidelines will be presented at the AAA's annual meeting in November.

Animal-rights protesters target drug companies

London Police arrested 87 animal-rights campaigners following weekend demonstrations and break-ins at the UK offices and laboratories of five leading pharmaceutical companies.

Hundreds of protesters took part in the action, claiming that the companies support the testing of drugs on animals at the research organization Huntingdon Life Sciences.

The arrests were made after protesters

broke into the Buckinghamshire facility of German chemicals and drugs giant Bayer, smashing windows and machinery. Other protesters broke into Eli Lilly in Hampshire and GlaxoSmithKline in Surrey and Berkshire. The campaigners claimed they also demonstrated outside the homes of directors of Roche and Pharmacia.

UK names sites for transgenic crop trials

London The British government has released details of the next stage of its genetically modified crop field trials, setting off a now-familiar cycle of sometimes ill-tempered debate and demonstration.

A total of 96 new trial sites of genetically modified maize, oilseed rape and beet will be sown in the coming months, doubling the size of the current test programme. The buffer distance between the transgenic and conventional crops, intended to stop cross-pollination, will be increased — from 50 to 100 metres for oilseed rape and from 50 to 80 metres for maize.

Organic farming and environmental groups have reacted angrily to the announcement, branding the separation distances as “pathetically inadequate”.

Indian earthquake raised underground rivers

New Delhi The 26 January earthquake that killed thousands in Gujarat state brought buried water channels to the surface of the drought-ridden Rann of Kutchh district, researchers report.

The new water sources were revealed by images from Indian remote-sensing satellites that pass over the country every five days. “There is no doubt that the signatures we see are those of water,” says Shailesh Nayak, head of the Marine and Water Resources Division of the Space Applications Centre in Ahmedabad. “What we are not sure of is whether these sources are permanent or short-lived.” Teams of scientists have been dispatched for on-site investigations.

Janardhan Negi, a senior scientist at the National Geophysical Research Institute in Hyderabad, says the emergence of old river channels is not surprising, as the area used to be a delta of the River Indus before it was submerged during an earthquake in 1819.

Lake Nyos surrenders its carbon dioxide

London An expedition to tame a ‘killer lake’ in northwest Cameroon was a success, according to members of the international team who



Hidden depths: the Cameroon lake harbours large quantities of dissolved carbon dioxide.

returned from the country last week.

The group of scientists went to Cameroon to install a pipe in Lake Nyos to remove carbon dioxide gas dissolved in the lake’s water (see *Nature* **409**, 554–555; 2001). An explosive release of carbon dioxide from the lake killed nearly 1,800 people in 1986.

There were concerns that the team might trigger another release of the gas as they sank 200 metres of plastic pipe into the lake and began siphoning off water containing nearly three times more dissolved carbon dioxide per unit volume than is found in a bottle of champagne.

As the team prepared to leave Nyos, the pipe was removing more carbon dioxide from the lake than was entering through the soda springs that feed it, team members say.

Back in business

The focus of activity in high-energy physics is about to switch from CERN, near Geneva, to Fermilab in Illinois. Sarah Tomlin sampled the atmosphere, as excited physicists prepared their Tevatron accelerator for action.

On a bright January afternoon at the Fermi National Accelerator Laboratory near Chicago, more than 200 physicists are gathered in 'the pit' — a huge chamber 15 metres underground. They are part of a 500-strong team that has, over the past five years, lovingly rebuilt DZero, one of two giant detectors at the Tevatron, Fermilab's main particle-accelerator.

The physicists are gathered for a photocall, and there is a buzz in the air. The 5,000-tonne detector, some three storeys high, is ready to be rolled into the path of the Tevatron's particle beams. And on 1 March, when the Tevatron is switched on, DZero and its sister detector, CDF, will become the centre of the universe for high-energy physics — a position they should occupy for at least five or six years. "This is where there's going to be really big discoveries made," enthuses Joe Lykken, a theorist at Fermilab. "You can have surprises in other kinds of experiments, but there's no substitute for probing the high-energy frontier."

Fundamental frisson

Under any circumstances, the weeks preceding the reopening of the world's most powerful accelerator would be an exhilarating time. But events last autumn at CERN, the European Laboratory for Particle Physics near Geneva, have provided an additional *frisson*. CERN's Large Electron-Positron (LEP) collider recorded four events that gave tantalizing evidence for the most elusive of subatomic particles, the Higgs boson. With a mass of 115 GeV, or 115 times the mass of a proton, the Higgs is thought to give other particles their mass. And it is the next major quarry for high-energy physicists. LEP is



Smile please: the DZero team strikes a pose.

now being dismantled, having missed its chance to secure the discovery. And that leaves the field open for the Tevatron.

Then there is Fermilab's role as the standard-bearer of US high-energy physics. Since 1993, when Congress cancelled the Superconducting Supercollider (SSC) project, the future of high-energy physics has been tied closely to CERN and its Large Hadron Collider (LHC), which should start running in 2006. But the revamped Tevatron will provide a brief window during which Fermilab will be the centre of attention. And that could prove crucial, as US physicists try to convince their political paymasters that the next big machine, beyond the LHC, should be built in the United States. "Fermilab is an ideal site for any of the future machines," argues Michael Witherell, the centre's director.

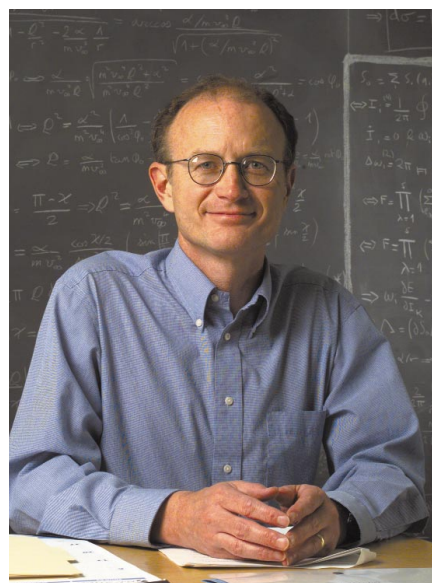
Gold standard

For the past 30 years, the theory of the basic constituents of nature — the Standard Model of particle physics — has been rigorously tested in accelerators around the world. And physicists have now discovered the full set of elementary particles predicted by the Standard Model.

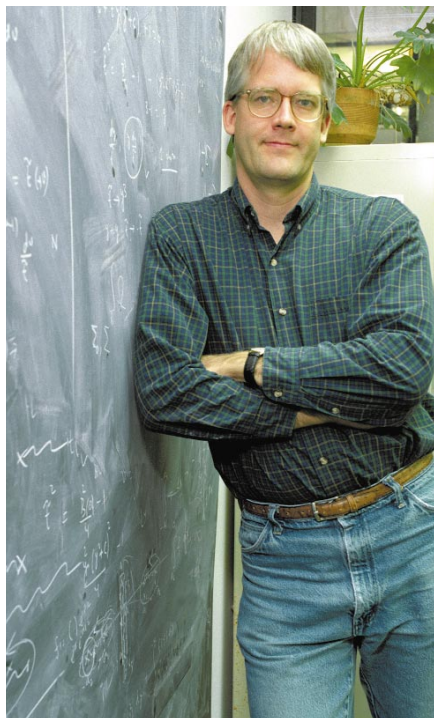
There are two basic kinds of particles, building blocks of matter (fermions) and carriers of a fundamental force (bosons). The last piece of the puzzle, the tau neutrino, was discovered last summer at Fermilab, by physicists studying data from Tevatron experiments done in 1997, in which beams of high-energy protons were fired at a tungsten target¹. This followed the Tevatron's previous triumph, the detection in 1995 of the freakishly massive top quark^{2,3} — which has almost the same mass as a gold atom.

Despite these successes, the Standard Model is actually little more than a recipe. Although physicists have a list of the Universe's fundamental ingredients, no one can explain why they are present in the quantities we observe. Furthermore, a little extra bit of magic is needed to give the recipe substance. This is the Higgs field, which permeates everything, giving particles their mass. Detecting its signature particle — the Higgs boson — is the closest physicists can get to proving that the Higgs field exists. "In some sense, the top quark and the tau neutrino are the end of a generation of discoveries," says Witherell. "And the Higgs is really the first of the next."

Fermilab's physicists will hunt for the



Michael Witherell: the hunt is on for the Higgs.



Joe Lykken: expects great discoveries.

Higgs in the debris that results when protons and antiprotons, both moving at close to the speed of light, collide head on. Since the end of the last collider experiments at Fermilab in 1996, a new \$250-million main injector has been built, increasing the intensity of these beams by a factor of 20. The detectors have had a \$200-million makeover, so that they can handle up to 10 million collisions per second.

Up close, CDF and DZero are astounding. Stand inside one of these behemoths and you can only marvel at their complexity. Each detector has a million electronic elements that can track the positions of particles. More than 1,200 kilometres of cabling carry away electronic signals, or feed cooling liquids or tracer gases, into the detectors. Through the middle of all this runs the Tevatron's 25-mm beam pipe.

Analysing data from the proton-antiproton collisions for signs of the Higgs will be more challenging than the task that faced physicists at LEP. The electrons and positrons smashed together at LEP are elementary particles, but the Tevatron's protons and antiprotons are effectively bags of quarks and antiquarks. When they collide, the result is a lot messier and harder to interpret. "You could get a Higgs boson the first day you turn on the accelerator, but to really know that you have a discovery takes years," warns Lykken.

The Tevatron will generate more collisions than can physically be recorded on tape. So out of the millions of collisions each second, the physicists plan to store only the 50 most interesting. The decisions on which ones to record will be made by 'triggers' — three levels of electronics and software systems that will sort out rare sightings from

humdrum events. The bottom quark, in particular, is a strong indicator of new physics. "It's crucial to Higgs searches," says Witherell.

Al Goshaw, co-spokesman for the CDF experiment, hopes the team will have its first serious results by 2002. But no one is expecting the Higgs to reveal itself that quickly — indeed, its discovery may have to wait for another upgrade, in which the Tevatron's beam intensity will be boosted a further five- or ten-fold, planned for 2004.

Desperately seeking SUSY

Despite the kudos surrounding the Higgs, many of Fermilab's physicists are even more excited by the possibility of the Tevatron uncovering evidence for an exotic theory that lies beyond the Standard Model. "A lot of theoretical work shows that the Standard Model is a patched together, ugly theory," says Lykken. According to this view, it may actually be embedded in a more elegant theory called supersymmetry (SUSY).

If SUSY is real, then every particle in the Standard Model also has a supersymmetric partner — the selectron for the electron, the photino for the photon, and so on. SUSY is attractive because it provides a way to unite fermions and bosons, which differ in a property called 'spin'. Some of its predicted particles should have masses comparable to the top quark, and might be within the Tevatron's reach. SUSY also predicts several 'flavours' of Higgs boson, each with a different mass.

But even stranger possibilities exist. Some physicists hope that the Tevatron will yield evidence for extra spatial dimensions. As many as 11 extra dimensions are required by string

theory, which is a contender for a deeper theory unifying the puny force of gravity with the other forces of the Standard Model. String theorists argue that we do not see these extra dimensions in our regular three-dimensional world because they are all rolled up very small.

Extra dimensions would probably show up in the Tevatron through their interaction with gravity. Like other forces, gravity is thought to be carried by a particle, yet to be detected, called a graviton. But with extra dimensions, huge numbers of extra gravitons could be created in high-energy collisions⁴, which may reveal themselves through their collective effects. For example, they might show up as missing energy in a collision, if a particularly heavy graviton disappears into another dimension.

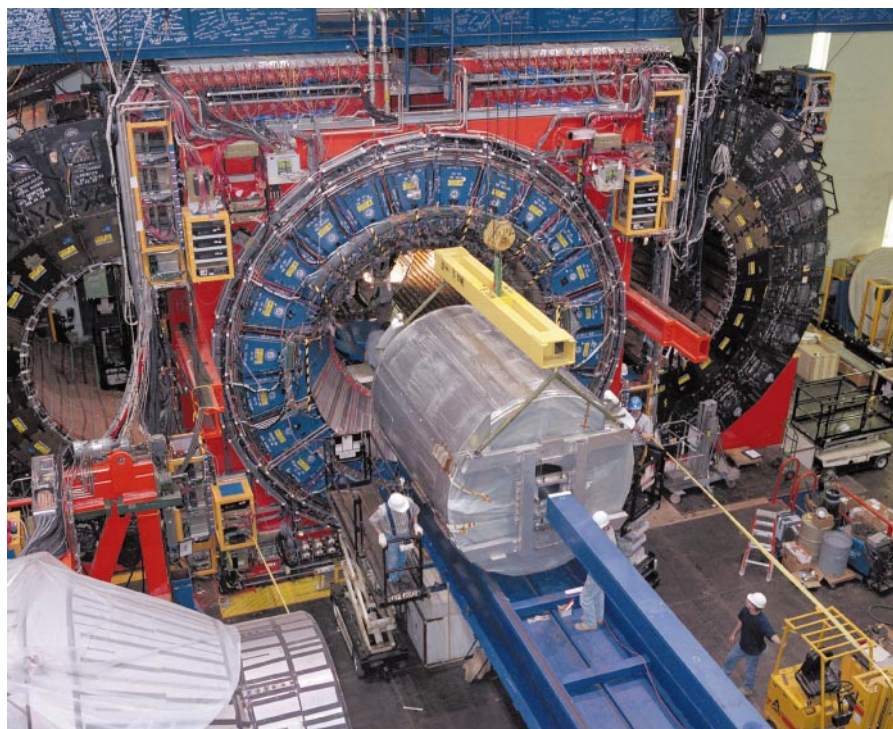
In short, no one can be quite sure what the Tevatron will discover. "It's not guaranteed that we will find any of these things," says Lykken. "But boy, it would be very disappointing if it didn't find at least one of them."

Amid the excitement and anticipation, however, there is one worry. According to one physicist, the LHC's start date hangs over Fermilab "like the sword of Damocles". With funding for the 2004 accelerator upgrade yet to be secured, there are still fears that the Tevatron could miss its chance to bag the Higgs. The clock is already ticking. ■

Sarah Tomlin is associate News & Views editor at Nature.

1. DONUT Collaboration *Phys. Lett. B* (in the press).
2. Abe, F. *et al. Phys. Rev. Lett.* **74**, 2626–2631 (1995).
3. Abachi, S. *et al. Phys. Rev. Lett.* **74**, 2632–2637 (1995).
4. Arkani-Hamed, N., Dimopoulos, S. & Dvali, G. *Phys. Lett. B* **429**, 263–272 (1998).

♦ <http://www.fnal.gov>



Detective agent: the CDF central tracker is installed following its multi-million-dollar face-lift.

What a long, strange trip it's been...

1985

In 1985, Charles DeLisi, then associate director for health and environmental research at the Department of Energy (DoE), begins to discuss a mammoth project — of a scale unprecedented in biology — to sequence the complete human genome. DoE funding begins in 1987.



PHOTOGRAPHY ON THIS PAGE:
PETER MENZEL/SPL
STEVE MUREZ/RAPHO/NETWORK
CHRISTIAN VIOUJARD/FRANK SPOONER
RAPHAEL GAILLARDE/FRANK SPOONER
ROBERT HARDING
J. BERGER/T. LAUX/E. MEYEROWITZ

1988

The National Institutes of Health (NIH) establishes the Office of Human Genome Research in September 1988. Renamed the National Center for Human Genome Research (NCHGR) a year later, its director is James Watson, co-discoverer of the double helix structure of DNA. Watson's testimony to the US Congress, in which he pledged to devote a small fraction of the project's budget to 'ethical, legal and social' issues, had proved instrumental in garnering political support.



Early 1990s

With sequencing still slow and expensive, the genome project adopts a 'map-first, sequence-later' strategy. In the early 1990s, two Parisian laboratories, the Centre d'Etude du Polymorphisme Humain and Génethon, have an integral role in mapping — underlining the project's international character. The labs' driving forces are Daniel Cohen (top) and Jean Weissenbach. Later, the genome project constructs a higher-resolution map that is used to sequence and assemble the human genome.



1998



In May 1998, Venter forms a company to sequence the human genome within three years. The company, later named Celera, will use an ambitious 'whole genome shotgun' method, which involves assembling the genome without using maps. But its data release policy will not follow the Bermuda principles.

1999

The public project responds to Venter's challenge. By early 1999, it is on track to produce a draft genome sequence by 2000. Increasingly, the bulk of the sequencing takes place in five huge centres: at the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts; the Sanger Centre near Cambridge, UK; Baylor College of Medicine in Houston; Washington University in St Louis; and the DoE's Joint Genome Institute (JGI) in Walnut Creek, California. The centres' leaders are dubbed the 'G5'. Here, Robert Waterston of Washington University in St Louis and John Sulston of the Sanger Centre are pictured in a rare moment of relaxation, while Trevor Hawkins and Elbert Branscomb of the JGI prepare samples.



Super models

The complete genome sequences of model organisms are proving immensely valuable to biologists working on these species, and will also help interpret the human genome sequence. Published highlights to date include the yeast *Saccharomyces cerevisiae* (May 1997), the nematode *Caenorhabditis elegans* (December 1998), the fruitfly *Drosophila melanogaster* (March 2000, right), and the plant *Arabidopsis thaliana* (December 2000, left).



The draft human genome sequence published in *Nature* this week is the culmination of 15 years of work, involving 20 sequencing centres in six countries. Here, we present a reminder of some of the key moments.

1992

Francis Collins of the University of Michigan replaces Watson as head of NCHGR in April 1992. Watson had earlier clashed with Craig Venter, then at NIH, over the patenting of DNA fragments known as expressed sequence tags.



1992

Later that year, Venter sets up The Institute for Genomic Research (TIGR) in Rockville, Maryland. TIGR later sequences a host of bacterial genomes, starting with *Haemophilus influenzae*.



1996



In February 1996, at a meeting in Bermuda, international partners in the genome project agree to formalize the conditions of data access, including release of sequence data into public databases within 24 hours. These came to be known as the 'Bermuda principles'.

PHOTOGRAPHY ON THIS PAGE:
LIAISON
NIKKAN KOGYO SHIMBUN
JAMES KING-HOLMES/SPL

1999-2000



The first complete human chromosome sequence — number 22 — is published in December 1999. Chromosome 21 follows in May 2000, a collaborative effort led by German and Japanese groups under the direction of André Rosenthal and Yoshiyuki Sakaki, respectively. Sakaki (centre) is pictured here at *Nature's* chromosome 21 press conference in Tokyo.

2000



Outside, celebrations continue with Eric Lander of the Whitehead Institute, Baylor's Richard Gibbs, and Waterston and Richard Wilson from Washington University.



On 26 June 2000, leaders of the public project and Celera announce completion of a working draft of the human genome sequence. Collins and Venter are seen here on television with Ari Patrinos of the DoE, who cut through the animosity between the rival projects to broker the joint announcement at the White House in Washington.

This week



Finally, this week sees the publication of the draft genome, the public sequence in *Nature*, Celera's in *Science*.



Technical gurus

Without advances in sequencing technology, we would still be waiting to unveil our genetic blueprint. Double Nobel laureate Fred Sanger (pictured) of the Laboratory of Molecular Biology in Cambridge invented the basic technique of gene sequencing back in the 1970s. In the 1980s, Leroy Hood, then at the California Institute of Technology in Pasadena, introduced the first automated sequencing machine. But it was the ABI PRISM 3700 DNA Analyzer, developed by Michael Hunkapiller of PE Biosystems, which allowed the rapid sequencing progress made by both Celera and the public project over the past two years. Assembling fragments of the genome into a complete sequence, meanwhile, depended heavily on computer programs developed by Philip Green of the University of Washington in Seattle.

We are grateful for contributions and input from Francis Collins, Richard Gibbs, Victor McKusick, John McPherson, David Stewart and the staff of the Cold Spring Harbor Laboratory.

Are you ready for the revolution?

If biologists do not adapt to the powerful computational tools needed to exploit huge data sets, says Declan Butler, they could find themselves floundering in the wake of advances in genomics.

“We’re in the middle of a messenger RNA in a gene here, TCRA, apparently. Before, we were at the base-pair level, now we have zoomed out to 53,000 bases. Let’s zoom out again, click on the 10×-out button — now we are at half a million bases.” David Hausler, a computer scientist at the University of California at Santa Cruz, is giving me a guided online tour from my workstation in Paris of a stretch of chromosome 14 using a web-based human-genome browser.

A few minutes into the demonstration, Hausler heats up. “Hey! Look over there at that SNP track: I see one right in the middle of an exon at the bottom of the screen; there is a line-up of all those bars. I’m going to click on that to see what we get. OK, it says it’s RS1197. That’s a hell of an interesting SNP to look at. Hey, it’s fascinating! Every time I demo this, we always discover something new and interesting! This means it’s a coding SNP.”

If you are one of the many biologists for whom genome databases are as comprehensible as a mass of supermarket barcodes, Hausler’s enthusiasm probably leaves you cold. If so, the publication this week in *Nature* and *Science* of draft human genome sequences marks a good time to team up with a friendly bioinformaticist and join the action, before it is too late.

Brave new world

Many research leaders predict that the potential to integrate different levels of genomic data — such as raw sequence from the human genome and those of model organisms, data on genetic variability between individuals and on gene expression in different tissues — will radically change biological research. They argue that small experiments driven by individual investigators will give way to a world in which multi-disciplinary teams, sharing huge online data sets, emerge as the key players. Some foresee an era of ‘systems biology’, in which the ability to create mathematical models describing the function of networks of genes and proteins is just as important as traditional lab skills.

If so, those who learn to conduct high-throughput genomic analyses, and who can master the computational tools needed to exploit biological databases, will have an



Shape of things to come: large data sets arising from genome projects demand new skills of biologists.

enormous competitive advantage. Some experts even predict that the outcome of this natural selection will see many current top scientists, research groups and even whole institutes relegated to the second division.

“Every institution that expects to be competitive in this new era will need to have strengths in high-throughput genomic analyses and computational approaches to biology,” says Francis Collins, director of the National Human Genome Research Institute in Bethesda, Maryland.

Søren Brunak, director of the Center for Biological Sequence Analysis at the Technical University of Denmark in Lyngby, is more blunt: “Biologists for the past 20 years have been of reasonable quality, but now we will see those who are good and those who are bad. Either you use these new methods in all aspects of your work, or you go on at your old speed and get left behind.”

Given such gung-ho predictions, many researchers are becoming concerned about the fate of traditional ‘wet-lab’ biologists. According to David Jones of the Protein Bioinformatics Group at Britain’s University of Warwick, some biologists are already having their grant proposals turned down

because they lack a bioinformatics component. “This is certainly going to become more common, and for a while there will be some gnashing of teeth as biologists take stock of this,” he says.

Ewan Birney of the European Bioinformatics Institute (EBI) in Hinxton, near Cambridge, identifies the need to transfer the skills of his discipline to the wider biological community as the single biggest challenge ahead. He fears that many biologists risk being “disenfranchised”. “Even a biologist doing very directed experiments is not going to be able to avoid large-scale data gathering and analysis,” Birney predicts.

Gene surfing

The simplest program in the bioinformaticist’s tool-box is known as BLAST (basic local alignment search tool). This allows a biologist who has identified an interesting gene sequence to compare it against those held in genome and protein databases in order to check, for example, whether it corresponds to a gene that has already been characterized. Then there are computational tools such as Genscan and GeneWise, which predict the occurrence of genes with-



Petrified? Lab-based researchers will need to swallow their fears and embrace computational biology.

in raw sequence data. And genome browser programs allow users to see whole chromosomes and to zoom in on regions of interest, while simultaneously superimposing associated information such as data on homologous genes from other species.

Now tools are being developed by several groups to analyse entire genomes and biosynthetic pathways, and to integrate multiple levels of information into mathematical models of cells and even organisms — introducing to biology the high-end computing previously restricted to fields such as astrophysics and climate modelling.

So far, expertise in using these tools has tended to reside in the main genome centres, or in specialist bioinformatics institutes such as the EBI or its US counterpart, the National Center for Biotechnology Information in Bethesda. Leroy Hood, director of the Institute for Systems Biology in Seattle, describes this separation between bioinformaticists and experimental biologists as a “fatal flaw”. “I would argue very strongly that you have to integrate very tightly the bioinformatics with doing experiments,” he says.

Birney accepts that people like him have to redouble their efforts to reach out to

experimental biologists. “Bioinformaticists must become very focused on being a service community,” he says. “We have a duty to make the data more usable in simpler ways through the web.”

In the meantime, many biologists who are trying to get up to speed with bioinformatics are using only the simplest tools available. Frequently, says Teresa Attwood, an expert in protein sequence analysis at the University of Manchester, they fall into “the black box trap”, accepting without question the matches returned by BLAST without taking the time to learn the caveats of such techniques.

Similarly, computational tools used to assign function to sequences have inherent limitations; put them in the hands of someone who is unaware of this and you have the makings of havoc. As a result, the public genome databases are already littered with annotations that are inaccurate, misleading or downright wrong.

Dig the new breed

So if the majority of biologists are not to be disenfranchised, and the databases are not to become terminally polluted with misleading information, what is the solution?

In the long run, change will come through the emergence of a new breed of biologists who are steeped in computational biology as an integral part of their education. This means that the subject must be included as a core module in all undergraduate biology courses, rather than as a specialist option. Although this is starting to happen, the availability of teachers with the appropriate expertise is still a limiting factor.

Funding agencies are also trying to drive change by ploughing money into initiatives that require a multidisciplinary approach and a strong computational component. The US National Institutes of Health, for instance, through its National Institute of General Medical Sciences in Bethesda, has created a programme of ‘glue grants’ for integrative and collaborative approaches to research. Under this programme, the Alliance for Cellular Signaling, headed by Al Gilman at the University of Texas Southwestern Medical Center at Dallas, will receive up to \$25 million over the next five years to draw up a complete map of the interactions between some 1,000 proteins in two types of cell. The consortium unites traditional experimentalists with computational biologists, with its membership drawn from institutions across the United States. “The glue grants is an exemplar flagship initiative that is refocusing the approach to biology research,” says Shankar Subramaniam, a bioinformaticist at the San Diego Supercomputing Center, based at the University of California at San Diego.

Some bioinformaticists are also working on ‘outreach’ activities. Birney, for instance, heads Ensembl, a joint project between the EBI and the Sanger Centre, also in Hinxton. This seeks to simplify things for mainstream biologists by making annotated genome sequences available through an intuitive web-based interface. Others, such as Subramaniam, are designing user-friendly



Ewan Birney fears that many biologists will be disenfranchised by the changes occurring.

▶ bioinformatics 'workbenches', which combine various computational tools.

Hood, meanwhile, hopes to drive the ethos of systems biology into the wider biological community. He plans to invite talented biologists to Seattle to take a crash course in working on large-scale data sets, so that they can then go back to their labs to use the data to pursue hypothesis-driven research. "Say you have a top-flight scientist working on the *p53* tumour-suppressor gene in Paris," says Hood. "We would bring him out; show him the kind of tools we have here, show him how you can interrogate whole systems at a time, and show him the utter importance of being able to integrate data from these different levels."

Hood's crash course may sound a little daunting. But if you decide to enter the world of bioinformatics, and find an expert who is willing to share his or her skills, it may prove easier than you think. The major problem, says David Roos, director of the Genomics Institute at the University of Pennsylvania in Philadelphia, is "lack of familiarity — not even knowing where to begin — and lack of people to provide that familiarity". Birney's advice to scientists debuting in bioinformatics is disarmingly simple: "Don't worry if you feel like an idiot, because everyone does when they first start."

In this regard, the experience of some of the biologists commissioned by *Nature* to mine the human genome for interesting nuggets of data is instructive (see pages 827–859). "The more one works, the more confident one becomes," says Rossella Tupler of the University of Massachusetts Medical School in Worcester, who searched for genes involved in processes including transcription and RNA splicing (see pages 832–833). "During the analysis for *Nature*, I increased my ability to analyse the data. I also felt very happy about the results of my search. We found several new genes."

Apocalypse postponed?

Given experiences such as these, some experts argue that the apocalyptic visions of thousands of top-notch biologists being suddenly cast aside by the genomic revolution are wide of the mark. "Molecular biology has smoothly absorbed any number of technological 'revolutions' as fast as the next generation can learn the new techniques," says Sean Eddy, a self-taught computational biologist at Washington University in St Louis. "Genomics and computational biology won't be any different. Molecular cloning, DNA sequencing and the polymerase chain reaction are all examples of revolutionary techniques. I bet you can dig up old quotes from the 1970s of people saying, 'Boy, we're all going to lose our jobs now. How will we ever retrain all the old guard to clone and sequence?'"

Stan Fields, a geneticist at the University of Washington in Seattle, adds that biologists



Easy does it: Shankar Subramaniam aims to make computational tools biologist-friendly.

studying the yeast *Saccharomyces cerevisiae*, which had its complete genome published in 1997, seem to be rising to the challenge. "Here, researchers from a predominantly genetic, small-laboratory approach have been confronted with the leading edge of genomic strategies," he says. "I think that this community has been highly successful."

When push comes to shove, says Gene Myers, vice-president for research informatics at Celera Genomics in Rockville, Maryland, molecular biologists will acquire skills in bioinformatics out of sheer necessity. "As traditional biologists see their colleagues outpace them because of technology, they will change," he says.

Nevertheless, Sylvia Spengler, a programme director at the US National Science Foundation's Division of Biological Infrastructure, believes that most biologists will need help to exploit large-scale data sets at the levels of sophistication needed. "You may not know how to use an integrated suite of tools," she says.

Among officials such as Spengler, one major concern is whether universities are geared up to provide the environment needed for bioinformatics and systems biology to flourish. This will require a truly multidisciplinary approach, cutting across departmental boundaries and embracing partnership, rather than competition, between major research universities in providing the necessary infrastructure. Many US universities have launched bioinformatics programmes and initiatives in computational biology, says Spengler. "But they are tiny, and they are very inward looking. It is not clear to me that any programme believes its mandate is to serve the entire community," she says. "Yale thinks it should serve the Yale community, Rockefeller thinks it should serve Rockefeller. But the picture is bigger."

Hood, who had to leave the University of Washington to get his Institute for Systems Biology off the ground, is more disparaging.

"The leadership in academia is utterly lacking, and their scientists are going to be at an enormous competitive disadvantage if they don't see the revolution that is coming," he says. "The new biology requires teamwork and putting together different kinds of science, and the current model of tenure is very much against that; in the United States it is based on what have you done yourself."

All systems go

Similar concerns are being expressed on the other side of the Atlantic. "Increasingly, university researchers are moving further away from the 'coal face'," says Warwick's Jones. "Without some radical changes in the types of research, and the levels of funding support, university biology departments are in danger of becoming spectators in genomics." Some scientists fear that the action may become concentrated in multidisciplinary institutes with the vision to embrace systems biology, and in industry — which has wasted no time in getting up to speed with computational biology.

But once again there are some who feel that these visions are the product of excessive revolutionary zeal — or are at least premature. "The bulk of contemporary molecular, cellular and model-organism biology is a tremendously successful enterprise and is doing fine without us," says Roger Brent, associate director of research at the Molecular Sciences Institute in Berkeley, California, and a leading figure in systems biology.

Tom Pollard of the Salk Institute for Biological Studies in La Jolla, California, who was commissioned by *Nature* to mine the genome for information relating to the cytoskeleton (see pages 842–843), agrees. "I have a personal fear that full genome sequences will prove irresistibly seductive, perhaps even deferring the hard wet-lab work required to complete the mechanistic analysis needed to understand physiology," he says. "Understanding function requires lots of hard-won information that can only emerge from analysis of one or a few types of molecules in a single experiment."

But what seems clear is that the research teams that will be most successful in the coming decades will be those that can switch effortlessly between the lab bench and a suite of sophisticated computational tools. If you are not already *au fait* with computational biology as you are with gels and culture dishes, you have been warned.

Declan Butler is *Nature's* European correspondent.

University of California at Santa Cruz Genome Browser

♦ <http://genome.ucsc.edu>

European Bioinformatics Institute

♦ <http://www.ebi.ac.uk>

National Center for Biotechnology Information

♦ <http://www.ncbi.nlm.nih.gov>

Ensembl

♦ <http://www.ensembl.org>

San Diego Supercomputer Center Biology WorkBench

♦ <http://workbench.sdsc.edu>

Skilled eyes are needed to go on studying the richness of the soil

Sir—Many ecologists are unaware that most of the biodiversity in any terrestrial ecosystem occurs in the soil, so we were delighted to see your recent News Feature¹. We do not agree, however, with the implication that such studies have been initiated only recently: in fact, the biodiversity and functioning of soil organism communities have been a major research area for soil ecologists for 50 years.

As early as 1950, Kühnelt² stressed the diversity of soil dwellers. A wave of publications outlining soil organism diversity occurred in the 1970s, including the recognition of the 'enigma' of high species diversity associated with apparent high trophic overlap³. The importance of soil organisms for plant growth has also been recognized for more than a century. Currently, soil biodiversity is even being seen as a tool for sustainable agriculture⁴.

Progress in soil ecological research has often been constrained by taxonomic and methodological problems, which are not resolved in some of the studies cited in your News Feature, making it difficult to reach meaningful levels of taxonomic resolution. Experimental manipulations such as removing starfish from a stretch of rocky shore are said to be extremely difficult to apply to soil-dwellers. Yet manipulative studies have been carried out with soil organisms for many years. These involve many different types of experimental approach including the use of litter-bags with different mesh sizes, partial or complete defaunation using pesticides, transplantation of soils and inoculation of experimental soil units. These experiments achieved in the field during the past 20 years or more, however difficult they are to perform, have been instructive in supplementing laboratory results and providing foundations for the studies cited¹.

The canopy has been called the 'last biotic frontier of the biosphere'⁵, and perhaps the soil can join it⁶. Unfortunately, funds for the study of nematodes, mites and other minute soil animals disappear into the ravening maw of molecular biology. If there is no one left who can tell one soil animal from another, however, we will be exploring these frontiers blind.

Henri M. André*†, **Xavier Ducarme†‡**, **Jo M. Anderson§**, **David A. Crossley Jr||**, **Hartmut H. Koehler¶**, **Maurizio G. Paoletti#**, **David E. Walter&**, **Philippe Lebrun†**

*UR Faune du sol, Musée royal de l'Afrique centrale, B-3080 Tervuren, Belgium

†Unité d'Écologie et de Biogéographie, Université catholique de Louvain, Place Croix du Sud 4/5,

B-1348 Louvain-la-Neuve, Belgium

‡Aspirant au Fonds national de la Recherche scientifique, Belgium

§School of Biological Sciences, University of Exeter, Exeter EX4 4PS, UK

||Institute of Ecology, Ecology Annex, University of Georgia, Athens, Georgia 30602-2360, USA

¶Universität Bremen FB2, Institute of Ecology and Evolutionary Biology, Centre of Environmental Research & Environmental Technology, Leobenerstr., PO Box 330 440, D-28359 Bremen, Germany

#Dipartimento di Biologia, Università di Padova, via U. Bassi, 58/b, I-35121 Padova, Italy

&Department of Zoology & Entomology, Hartley-Teakle Building, University of Queensland, St Lucia, Queensland 4072, Australia

1. Copley, J. *Nature* **406**, 452–454 (2000).

2. Kühnelt, W. *Bodenbiologie* (Herold, Vienna, 1950).

3. Anderson, J. M. in *Progress in Soil Zoology* (ed. Vanek, J.) 51–58 (Academia, Prague, 1975).

4. Paoletti, M. G. *Agriculture, Ecosystems & Environment* **74**, 1–18 (1999).

5. Erwin, T. L. *Bull. Ent. Soc. Amer.* **29**, 14–19 (1983).

6. André, H. M., Noti, M. I. & Lebrun, P. *Biodivers. Conserv.* **3**, 45–56 (1994).

Astronomy network will allow every site to shine

Sir—The News Feature on optical telescopes (*Nature* **408**, 12; 2000) reviews the present state of most national telescope operators in Europe and the United States, and identifies some important niches that medium-sized (2–4-m class) telescopes will fill in the future. The story seems to suggest, however, that it is necessary to close most nationally operated 2–4-m-class telescopes in Europe and focus operation on a single site. This is not what has been discussed recently.

What is required is that national telescope operators establish a single European network to coordinate operations of the observatory facilities and the instrumentation of their telescopes. This will allow each site to focus on its particular strengths, while assuring a comprehensive instrumentation complement across the network. The coordination will introduce synergies and allow more efficient operation at each site, which in turn will reduce costs. Without such coordination among national facilities, individual telescopes might no longer achieve the funding to remain competitive and might close prematurely.

At the same time, the network should offer time to European countries that do not operate national observatories or which have no general access to international observatories such as the European Southern Observatory. Access to other European facilities may foster astronomy programmes in these countries, in the way that Spanish astronomy has benefited from

access to installations such as the Institute of Millimetric Radioastronomy in the French Alps, the Isaac Newton Group (ING) on La Palma in the Canary Islands and Calar Alto on mainland Spain.

There has been a very large response to the creation of such a network. During a recent meeting of the Optical and Infrared Coordination Network for Astronomy (OPTICON), members agreed to study a system of instrument coordination and development, and to consider systems to extend telescope access, exchange time between telescopes, and enhance the education and training role of its facilities.

The expected synergies are illustrated by ING and the Calar Alto observatory. With Omega2000, Calar Alto will be operating a near-infrared camera that provides the largest field of view at any telescope in the northern hemisphere. On La Palma, on the other hand, instrumentation efforts focus on adaptive optics and wide-field spectroscopy. Access to both wide-field imaging and wide-field spectroscopy at two different sites offers obvious benefits to the broader European community.

Therefore, medium-sized telescopes can have an exciting future. The undersigned participate as coordinators of two OPTICON working groups which aim to implement the ideas described here.

Roland Gredel*, **Rene Rutten†**

*Calar Alto Observatory, Jesus Durban 2,2,

Apartado de Correos 511, E-04080 Almería, Spain

†Isaac Newton Group of Telescopes, Apartado de Correos 321, 38700 Santa Cruz de La Palma, Islas Canarias, Spain

Semmelweis and the battle against infection

Sir—Fred Rosen's review of *The War Within Us* by Cedric Mims (*Nature* **408**, 769; 2000) mentions Semmelweis, who introduced antiseptics into medical practice. Born in Budapest in 1818 to a family of German descent, he was called Ignaz Philipp in German and Ignác Fülöp in Hungarian. He considered himself Hungarian and attended medical school in Hungary, but pursued some of his professional career in Vienna where he encountered both success in his work, and humiliation by his jealous boss.

Being male, however, he did not exactly die from "the very disease he spent his life trying to eradicate" — since this was puerperism, or childbed fever. Rather, he died of staphylococcal infection, the agent that is most often responsible for post-partum death.

George Rédei

3005 Woodbine Court, Columbia, Missouri 65203-0906, USA

Patents in a genetic age

The present patent system risks becoming a barrier to medical progress.

**Martin Bobrow
and Sandy Thomas**

In helping to develop many discoveries into clinically useful products, the patent system has been a force for good. It has encouraged innovation and sensible risk taking, stimulated investment in research and development, and delivered drugs and devices that have changed the face of modern medicine.

But can the patent system maintain this positive influence in the future? As we stand on the brink of one of biomedical science's greatest achievements — the complete sequencing of the human genome — this question becomes extremely pertinent. The issue of how and when to assign patent rights to all manner of DNA sequences threatens to throw the patent system into disarray.

The patenting system should help people to channel their energy towards inventions of genuine therapeutic or diagnostic value and discourage the frenetic cataloguing of DNA sequences that are a long way from being a final, useful product. But as it stands, the system is in danger of not fulfilling these functions. It can be slow, and prohibitively expensive to navigate. Moreover, patenting is increasingly being seen by some as rewarding luck rather than enterprise.

Retuning

The present state of affairs is partly a result of the way the patent system has evolved in recent years. It has changed from focusing on conventional drugs to being a system that also encompasses patents on biological molecules containing genetic information (see Box, overleaf). Many thousands of patents with claims to human DNA sequences have been filed and granted¹, and few have as yet been subject to legal challenge. The many patents include claims for genomic DNA sequences, complementary DNAs, individual mutations, expressed sequence tags (ESTs) and single nucleotide polymorphisms (SNPs; see glossary on page 815). Unless the system is retuned, there is a serious risk that complexity will heap upon complexity, making it unmanageable.

Will these patented sequences actually be used to develop new healthcare products, and if so, how? Some already have been — for example, in the gene-based diagnostic tests for cystic fibrosis and breast cancer. But the granting of exclusive property rights for disease genes so that they are under the sole control of the patent holder has already generated wide debate about whether claims



Debate continues over whether DNA sequences can be claimed to be part of an invention.

to DNA sequences as part of an invention can continue to be justified.

Many gene patents are also being filed for their potential, though unpredictable, role in drug development. But apart from in rare situations, it is generally the protein products, their derivatives and antagonists, rather than the encoding DNA, that will bring medical benefits. Proteins, and even parts of proteins — characteristic arrangements of molecules (motifs), folds and subsections — are already subject to the same welter of patent claim and counter-claim as DNA sequences. For example, one US patent² covers a computer-assisted method for identifying compounds possessing a similar structure to the hormone erythropoietin. Anyone using the three-dimensional coordinates to identify crystal structures similar to a specified peptide might infringe the patent³.

In the human body, the components of the functional genetic units — exons, SNPs, mutations, protein motifs, control regions and transcripts — are all integral parts of the same biological mechanism that interacts with other gene products. If the DNA sequences of all of these components are identified and then treated as separate 'inventions', any useful product is highly likely to cross the boundaries of several patents. The ensuing accumulation, or 'stacking', of patents⁴ on a product requiring licences would be lucrative for lawyers, and profitable

for the few people fortunate enough to have rights to the key elements of genetic information. But it is hard to see it as a defensible means of rapidly developing cost-effective and useful products.

Claims that such problems will not occur, because the patent system is self-correcting through legal challenges, do not stand up to scrutiny. A legal challenge is generally slow and expensive, and its outcome is unpredictable. Moreover, patent holders from academia and small companies are particularly disadvantaged. Long and expensive litigation, and conflicting judgements on what should be patented, are not a sensible way forward in such an important and fast-moving area. Furthermore, although many patent claims to human DNA sequence may well eventually fail under legal challenge, until they do they are an expensive barrier to progress in applying the fruits of medical research to clinical and therapeutic practice.

Policy failures

Who are the villains of the piece? Not the scientists, nor the individuals in business — they are playing within the rules. Nor the lawyers, who are appropriately pursuing their clients' interests. Nor is it the understaffed patent offices, which are under constant pressure from well-resourced proponents. We suggest that it is the lack of leadership from policy-makers that has allowed the present state of affairs to come about. Our elected representatives — put off, no doubt, by the complexities of the patent system, strong vested interests, and a reluctance to address the international dimension — have failed to develop public policies for biotechnology that address the weaknesses of the patent system yet retain its strengths.

In the absence of serious legislative action, policy has more or less evolved through dialogue within a limited circle of participants. Commercial interests, which are well represented to the patent offices, have not been counter-balanced by those who represent the broader public interest. The result has been an innate tendency for the patent system to 'creep' in the direction of extending patentability to biotechnology inventions for which the thresholds for novelty, inventiveness and utility have been lowered.

Legal challenges from other parties can redress the balance of interests, but this route is often eschewed by the private sector in favour of individual licensing deals — it is often quicker and less expensive to arrange a cross-licensing agreement than to challenge

a claim in court. And there has been remarkably little assessment by elected representatives of whether the system's current balance is likely to serve the public interest.

The fact that policy-makers have done little to reform the patent system is surprising given the importance of patenting in healthcare. Why is this? Policy has broadly evolved at four levels: through international agreements such as the World Trade Organization's Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS); through regional and national legislation such as the Directive on the Protection of Biotechnological Inventions from the European Union (EU); through the patent offices; and through the courts. The main failing of this set of institutional arrangements is that it has not produced a coherent policy framework. Instead, the attention of policy-makers has been uneven and there have been few attempts to grapple with the patent system as a whole.

TRIPS contains no explicit reference to genetic material. In the EU, policy has been dominated by the protracted passage of the directive; and the legislation, passed in 1998, did little more than formalize current practice. Although the Netherlands has since challenged the directive in the EU Court of Justice in Luxembourg on procedural grounds, it is unclear where this action will lead.

In the United States, industrial lobbies have been influential in Congress and the Senate in maintaining what is essentially a pro-patent, liberal policy-framework for biotechnology. But there have been some recent welcome adjustments. For example, in response to the debate about the patentability of ESTs, the US Patent and Trademark Office has tightened up its utility requirement⁵.

Targets for reform

What more could policy-makers be doing to address these patenting issues? Reforming the patent system would undoubtedly be politically complex, particularly if it included attempts to achieve international agreement. But that is no reason not to try. The legal and managerial complexities of the current framework risk slowing progress, which is not in the interests of the vast majority of the world's population.

As a first step, there needs to be a clear articulation and discussion of the key questions: whether patents on DNA sequences can continue to be justified in the context of current technology; whether such patents are actually necessary for successful innovation in healthcare; what the thresholds should be for novelty, inventiveness and utility; and what the duties of patent holders should be in licensing their inventions. There is also an urgent need to obtain evidence that might help in evaluating how current and future practices may affect healthcare inno-

To patent or not to patent

The general principle of the patent system is that it protects an invention from commercial competition. An inventor is granted a time-limited monopoly to exploit an invention if it is shown to be new, not obvious, with a useful application, and is disclosed in sufficient detail to allow verification. When applied, for example, to machine parts, this process is relatively straightforward. But when applied to conventional drug development, compromises are necessary. Although some drugs are genuine inventions, designed and synthesized by human ingenuity, others are based on naturally occurring, pharmacologically active molecules. It could then be claimed that these are discoveries rather than inventions. But it is widely accepted that this class of drugs also warrants patent protection because the steps involved in extracting, isolating and purifying these molecules from their natural state to one of pharmacological utility are sufficiently inventive to warrant a reward for success.

But future drug discovery, as well as diagnostics and novel therapies such as gene therapy, will increasingly depend on genomic information. To a degree, the patent system has been adapted to cover gene sequences and related matters. But although patent law now recognizes the patentability of DNA sequences, this may be one of those uncommon instances in which the law has lost touch with the way the real world works.

Although many patentable inventions might lie along the road towards developing useful products from the genetic information encoded in DNA sequences, the starting point — the normal or abnormal naturally occurring gene sequence — is the isolation of naturally occurring information. The argument, which has been widely used by patent lawyers, that DNA sequence identification is a form of purification "outside the body", analogous to the purification of naturally occurring pharmacological agents, appears specious. The DNA molecule is not (in this context) important as a substance. Its value resides in its information content — knowledge, not objects.

Similar arguments have been put forward for patenting complementary DNA sequences (cDNA): the 'natural' gene contains introns — sequences of DNA that are not translated into the final protein — and so synthesizing a cDNA is an 'alteration' of nature and as such a human invention. But this, too, seems intellectually trivial. Messenger RNA exists in nature, and cDNA is just a translation of this sequence. This is rather like saying that the same invention could be repatented if translated into a different language.

Finally, there should be an objective assessment of the efficiency of the current frameworks — including the patent offices, courts and legislative bodies — in determining policy.

Clearly, these issues need to be addressed in an international policy forum, perhaps within the ambit of bodies such as the World Intellectual Property Organization or the World Health Organization. Alternatively, a new body might be created. But for all the talk of international harmonization, patents remain essentially national in nature. The US, European and Japanese jurisdictions would have to be considered separately. It is already late in the day to be contemplating an international policy forum given the rapid growth in the fields of functional genomics, structural genomics and proteomics. To be of any use, such a forum must be able to influence policy-makers. And it should be set up without delay.

In arguing for change, we acknowledge that the patent system has, on the whole, worked well. But it has now moved into very different territory without appropriate policy guidance. The system does not need a complete overhaul. But it does need more than tinkering. In particular, it needs to adjust the stringency of the definitions of inventiveness and utility required for a patent⁶. Rigorous attention to such detail could greatly improve the system without altering its fundamental role in innovation.

We stand on the brink of biomedical science's greatest achievements. To make full use of them, we need the enthusiastic sup-

port and backing of the world's population. The grim signpost of genetically modified foods in Europe should remind us not to underestimate the power of public opinion. Epithets such as 'land grab' abound in press comments, on both sides of the Atlantic, on the recent spate of patents filed on parts of the human genome. The analogy is not misplaced. It has been a rush, and luck has often played a major part in determining which particular group gains the intellectual property rights over its competitors.

The human genome sequence is finite whereas human inventiveness is infinite. Public perception of scientists as being motivated by short-term financial gain is not only unfair to virtually all scientists, but also risks bringing biomedical science into public suspicion and disrepute, at a particularly sensitive time in its evolution. We can ill afford this risk.

Martin Bobrow is in the Department of Medical Genetics, Cambridge Institute for Medical Research, Wellcome/MRC Building, Addenbrooke's Hospital, Cambridge CB2 2XQ, UK.

Sandy Thomas is at the Nuffield Council on Bioethics, 28 Bedford Square, London WC1 3JS, UK.

1. <http://nytimes.com/library/national/science/062800sci-genome-patents.html>.
2. US Patent No. 5,835,382.
3. Meyers, T. C., Turano, T. A., Greenhalgh, D. A. & Waller, P. R. H. Patent protection for protein structures and databases. *Nature Struct. Biol.* 7 (Suppl.), 950–952 (2000).
4. Heller, M. A. & Eisenberg, R. S. Can patents deter innovation? The anticommons in biomedical research. *Science* 280, 698–701 (1998).
5. <http://www.uspto.gov>
6. Barton, J. Intellectual property rights: reforming the patent system. *Science* 287, 1933–1934 (2000).

The road to the code...

... and the cast who brought genetics to centre stage.

Cracking the Genome: Inside the Race to Unlock Human DNA/The Sequence: Inside the Race for the Human Genome

by Kevin Davies

Free Press/Weidenfeld & Nicolson: 2001.

288/320 pp. \$25/£20

Jan A. Witkowski

In 1901, William Bateson, the biologist who coined the word genetics, presented his first "Report to the Evolution Committee of the Royal Society". It included a footnote on "extraordinarily interesting" observations by Archibald Garrod, a physician at the Hospital for Sick Children in London: "Recently Garrod has noticed that no fewer than five families containing alkaptonuric members ... are the offspring of first cousins ... exactly the conditions most likely to enable a rare and usually recessive character to show itself."

A century after this first description of mendelian human genetics, *Nature* and *Science* contain the first reports of the 'complete' sequence of the human genome. This is a tremendous accomplishment, and even more remarkable when set in the context of those 100 years.

Human genetics was not an instant hit. Geneticists found fruitflies and other 'simpler' organisms more tractable material for analysis, and physicians had little interest in obscure disorders that affected so few people. (The eugenicists' interest in human genetics was another matter, being driven by social and political goals, rather than by scientific curiosity or medical necessity.) There were some early successes — notably the foundations of population genetics, laid by Ronald Fisher, J. B. S. Haldane and Sewall Wright (1908–50); the molecular nature of sickle-cell anaemia discovered by Linus Pauling (1949); and electrophoretic studies of inherited protein variations by Oliver Smithies (1955).

It was not until 1956 that the number of human chromosomes was determined correctly by Jo Hin Tjio and Albert Levan, and, some 20 years later, the first human gene (β -globin) was cloned. But modern human genetics did not begin until 1980, when David Botstein, Ray White, Mark Skolnick and Ron Davis showed how so-called restriction fragment length polymorphisms (RFLPs) could be used to find human disease genes.

Then, in the mid-1980s, discussions began on the desirability of sequencing the complete human genome. The project that was funded by the US Department of Energy began in 1987. It was followed in 1990 by that from the National Institutes of Health, under the directorship of James Watson.

the human genome project...



Hard sell: Jim Watson woos Leroy Hood, David Botstein, Sydney Brenner and Charles Cantor.

Robert Cooke-Deegan wrote a scholarly account of these early years of the Human Genome Project (HGP) in *The Gene Wars* (W. W. Norton, 1996). Now Kevin Davies' book *Cracking the Genome* covers the period from 1991 to the present. During those years, we were as likely to learn of the progress of the HGP from the pages of *The New York Times* as from those of *Nature* or *Science*. And if these accounts were to be believed, the twists and turns of the HGP were worthy of a Wagnerian epic.

The central characters in Davies' account are Craig Venter and Francis Collins. Venter founded the non-profit Institute for Genome Research before establishing the for-profit company Celera. His forte has been to see what needs to be done, and then to do it. Sydney Brenner and Ham Smith proposed very different ways in which the sequencing could be carried out, but it was Venter who acted decisively. Davies recounts how Venter and Smith began sequencing a microbial genome even before they had submitted a grant application to the National Institutes of Health. By the time the proposal was rejected as "impossible", the sequence was 90% complete.

Francis Collins, as the leader of the publicly funded sequencing efforts — Eric Lander calls them the "Forces of Good", leaving

us to supply the equivalent phrase for Celera — was constrained by his position from pushing ahead as decisively as Venter. The international battalions of the public genome effort, although fighting on the same side, did not always agree on how the war should be fought. There were also leaders in Congress to be cajoled, and here Collins was adept at keeping funds flowing to the project by persuading congressional committees of the benefits that would flow from it.

There is a large cast of characters, all equally worthy of taking centre stage. They include John Sulston and the Wellcome Trust in England. Both wielded tremendous influence — the former through his insistence on hewing to a gold standard of sequencing and the immediate public release of sequence data, and the latter through its financial clout and irrepressible envoy, Michael Morgan. At the 1998 genome meeting at Cold Spring Harbor Laboratory in New York, Morgan's announcement that the trust was to double its support of the Sanger Centre in Cambridge won a standing ovation from the assembled troops. The Sanger Centre is sequencing one-third of the human genome, a task paid for primarily by the Wellcome Trust.

Davies discusses the controversies over the patenting and commercial exploitation

book reviews

of the sequence. There continue to be irreconcilable differences between the demand by publicly funded laboratories for immediate and unrestricted release of data and the necessity for Venter to maintain some degree of secrecy in order to sell genomic information. The most recent manifestation of this controversy is the failure of the participants to agree on joint publication.

I am not sure who Davies sees as his audience. He includes a chapter on DNA that is aimed at non-scientists, yet much of the book requires familiarity with the genomics research world. Davies tries to show why sequencing the genome is both useful and fascinating by including chapters on the implications for clinical genetics, human evolution and gene therapy, but I found that these broke the flow of the central story of sequencing the genome. It could be argued that Davies' US title is not quite on the mark. So far, we have read the letters of the genome without much understanding; cracking the genome is yet to come.

Davies tells the story with verve, but tends to lapse into hyperbole, as if he feels the need to reassure the reader that this is a big story; for example, the double-helix paper stands "above the entire rest of the pantheon of scientific literature", and the completion of the human genome sequence is "perhaps the defining moment in the evolution of mankind". And there are mistakes; it was Robert Sinsheimer at the University of California at Santa Cruz, rather than Charles DeLisi of the Department of Energy, who first considered sequencing the human genome. Davies' claim that DNA was "virtually ignored" in the period between Friedrich Miescher (1871) and Oswald Avery (1944) displays a disappointing lack of knowledge of the studies of Albrecht Kossel, Phoebus Levene and William Astbury.

Davies closes the book with the charming story of Bateson's conversion to mendelism while travelling by train to give a talk at the Royal Horticultural Society. According to Bateson's wife, Bateson's reading of the experiments performed by the little-known monk 35 years previously transformed his view of heredity. Bateson revised his presentation en route to include Mendel's findings. Unfortunately, historian Bob Olby's recent research suggests that Bateson was reading, not Mendel, but the Dutch geneticist Hugo de Vries.

Davies did not set out to write a definitive record of the politics of the genome project, which is a pity; genome scientists are likely to be left wanting more behind-the-scenes details and scuttlebutt. *Cracking the Genome* is a good start, but the epic that is the Human Genome Project awaits its Boswell, or Watson. ■

Jan A. Witkowski is at the Banbury Center, Cold Spring Harbor Laboratory, PO Box 534, Cold Spring Harbor, New York 11724-0534, USA.



BETTMANN/CORBIS

Happy together? Einstein was said to have treated his first wife Mileva Marić with undisguised contempt.

Affairs of the heartless

Einstein in Love: A Scientific Romance

by Dennis Overbye
Viking: 2000. 416 pp. \$27.95

Lewis Pyenson

Seeking the laws of nature, physicist Albert Einstein invoked the Deity's love of harmony. For more than a decade, however, we have been aware that Einstein, the man, surrendered celestial harmony to earthly passion and engendered domestic dissonance. Einstein was the great defender of cause-and-effect in physics, but from his youth, he sought the affection of women, in parallel and in series, with little concern for the consequences. Celebrated as a gentle pacifist, Einstein treated his family with heartless disdain and his first wife with undisguised contempt. He raised large sums for charity while depriving his children of a tranquil existence. In *Einstein in Love*, science writer Dennis Overbye has provided an engrossing narrative of the personal and professional life of *Time* magazine's "Man of the Century".

Einstein's attraction to his first wife Mileva Marić, a free-thinking Serb of Austro-Hungarian nationality, echoed his sceptical attitude towards authority. Both were irreverent and unconventional students of physics. Early in their courtship he called her

Doxerl (Dollie) and she called him Johanzel (Johnny), perhaps allusions to characters in *A Doll's House* (Henrik Ibsen was a notable figure in Munich when Einstein was a boy there). Notwithstanding doubts expressed by the editors of the Einstein papers, there is circumstantial evidence that Marić shared in the genesis of relativity theory.

Just as Einstein took time to obtain a doctorate — failing several times in the attempt — he was also slow to reach maturity in personal relations. Shortly before he finally broke with Marić, when Europe was careening towards war in 1914, Einstein drew up a list of demands for keeping up appearances with her that reads uncannily as if it were the marriage contract of the emotionally crippled scholar Peter Kien in Elias Canetti's 1935 novel *Auto da Fé*.

When Einstein's station in life permitted it, he gave libido free rein. Women became, for him, an anodyne for the rigors of pure thought. Overbye portrays Einstein's second wife Elsa Einstein as a bourgeois comfort, suiting his station as a world-famous professor at Berlin University. Through her efforts, he ate well and recovered his Jewish heritage.

The charming affair between Einstein and Marie Curie, portrayed by Yahoo Serious in his underrated 1988 comedy film *Young Einstein*, is entirely fictitious, but it invites us to consider how Einstein loved. If we may infer from his correspondence, his love is largely sexless; momentous thoughts find little place in it. In addition to cute nothings, letters sent to his wives refer to

food and drink, to the small universe of an apartment, and to the recollection of a night of mutual comfort. There are petulant, uncharitable judgements along with naive self-deprecation. *Einstein in Love* reveals Einstein's sadness at his personal failings alongside his elation at the prospect of a tryst. We are a long way from John Donne's "Extasie", in which mutual attraction leads to transcendence.

To hear Einstein's physicist friends speak about him is to imagine a figure of great strength and sensitivity. Einstein's portrait in Leopold Infeld's 1941 autobiography *Quest* (Chelsea, 1980), for example, radiates humility, honesty and charity. In Infeld's account, and in many others, the man behind the intellectual revolution produced by the energy quantum and relativity has the face of a kindly, absent-minded sage whose own historical account of what happened (written with Infeld) has the comforting English title, *The Evolution of Physics*. It is hard to overestimate the impact of such an image.

In his account, Overbye juxtaposes Einstein's emotional state and his scientific research. Einstein's love affairs and his thoughts about physics appear serially without connecting commentary. (The physics receives unnecessary and sometimes confusing explanation, while there is little attempt to analyse matters of the heart.) The disjointed style reinforces the reader's sense that the qualities of kindness and humanity, as represented in the larger-than-life statue of Einstein on the Mall in Washington, are social constructions.

Einstein in Love ends abruptly with Einstein's triumph in predicting the deflection of starlight during the solar eclipse of 1919. The story carries a disturbing moral: repelled by the flawed genius of Wagner and Nietzsche, scientists in the twentieth century have portrayed this man as a personification of his great work. ■

Lewis Pyenson is at the Graduate School, University of Louisiana at Lafayette, PO Box 44610, Lafayette, Louisiana 70504-4610, USA.

An excavation of the drug myth

Intoxicating Minds

by Ciaran Regan

Weidenfeld & Nicolson: 2001. 164 pp. £14.99

Leslie Iversen

At first sight, Ciaran Regan seems to have attempted the impossible — a summary of the whole of psychopharmacology in less than 50,000 words aimed at a general readership. But he has been remarkably successful and has written an entertaining and informative account of mind-altering drugs, with a minimum of technical jargon or chemical structures.

The book covers essentially all of the drugs that are used recreationally — both the legal ones (caffeine, nicotine and alcohol) and the illegal ones (amphetamines, ecstasy, cocaine, cannabis, heroin and the hallucino-

gens). For all of these he manages to convey up-to-date information about how they work, at the same time giving the reader basic information about the brain and neurotransmitter mechanisms involved. The book gives a very good basic overview of brain function, including a digression into the mechanisms involved in memory.

The author takes a non-judgemental approach to his subject matter. As he puts it: "What is needed is an excavation of the drug myth. This [book] does not extol or condemn drug use; it simply invites reflection." This refusal to demonize drug-taking is refreshing.

Another large section of the book describes the drugs used to treat mental disorders — the anti-anxiety agents, antidepressants, antipsychotics and lithium. There are accurate and insightful descriptions of the illnesses themselves, with plenty of up-to-date information. The discussion of antidepressants, for example, accurately points out the irony that the pharmaceutical industry has now come full circle. The early antidepressant drugs imipramine and amitriptyline inhibited the reuptake by neurons of the two neurotransmitters noradrenaline and serotonin. The Prozac generation of serotonin-selective reuptake inhibitors replaced these. But the latest antidepressants are again combined noradrenaline- and serotonin-reuptake inhibitors — albeit with fewer side effects than imipramine and amitriptyline. Similarly, in the development of new drugs to treat schizophrenia there has been little success in breaking away from dopamine-receptor blockade as the universal mechanism of action.

The author writes in a lively and engaging style, and the text is full of anecdotes about how drugs were discovered or how their use originated, in many cases early in human history. I could find little to criticize in terms of the scientific accuracy of any part of the book, although the description of significant alcohol abuse as requiring the consumption of "at least a bottle of whisky a day for a period of several weeks" did seem to smack a little of Irish hyperbole!

Professional neuroscientists and pharmacologists will find little here that they do not already know — although they may find many of the historical anecdotes interesting and amusing. Did you know, for example, that the first antidepressant drug, the hydrazine derivative iproniazid, originated in part because Germany was left with a large stock of unused hydrazine rocket fuel at the end of the Second World War? The book can certainly be recommended to non-experts who want to know more about mind-altering drugs. ■

Leslie Iversen is in the Department of Pharmacology, University of Oxford, Oxford OX1 3QT, UK.



Whether we approve or disapprove, mind-altering drugs are central to the human experience.

One who created a tempo of his own

George Gaylord Simpson: Paleontologist and Evolutionist

by Leo F. Laporte
Columbia University Press: 2000. 314 pp.
\$50 (hbk), \$16 (pbk)

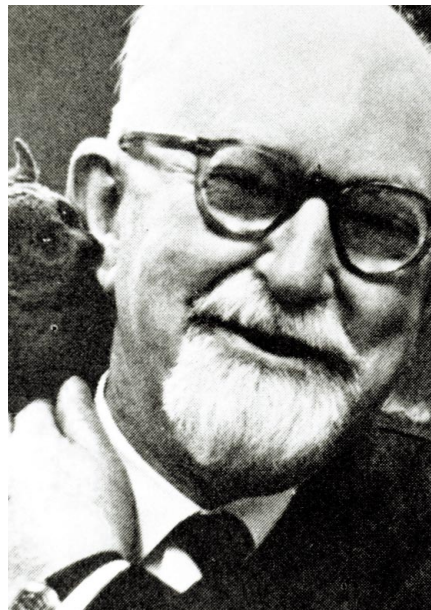
Michael J. Novacek

Palaeontology offers direct evidence about extinct life, and as such is integral to our understanding of nature and our place in it. But how exactly do patterns in the fossil record elucidate our understanding of evolution? Many decades ago, George Gaylord Simpson (1902–84), the subject of this informative and engaging biography by Leo Laporte, was consumed by this problem. In addressing those who refused to accept that palaeontological data could detect phylogenetic relationships and resolve other evolutionary problems, Simpson wrote: “In so difficult a study it is inexcusable to reject offhand any evidence that might give light.”

Laporte notes that Simpson “threw down the gauntlet” when he asserted that there was no reason why palaeontology and genetics should not be combined to provide a revised evolutionary perspective. He picked up the gauntlet himself in 1938 when he began writing *Tempo and Mode in Evolution*. With that work, Simpson, already known as an astoundingly prolific and influential palaeontologist, took on the mantle of visionary palaeontologist and, along with biologists such as Theodosius Dobzhansky and Ernst Mayr, became an architect of what was then called the “modern evolutionary synthesis”.

Laporte, whose book is a linked series of previously issued articles, treats the publication of *Tempo and Mode* in 1944 as a pivotal moment that transformed a career and a discipline. He effectively traces the influential roots of the work and methodically reviews its major tenets. He also reminds us that Simpson was the first to really codify the notion that microevolutionary processes suggested by genetics and population biology were essentially sufficient to explain patterns in the fossil record. Accordingly, Simpson thoroughly trounced vestiges of advocacy for ‘straight line’ evolution that seemed peculiarly enduring within palaeontology.

Laporte’s analysis of the impact of *Tempo and Mode* lacks incisiveness, and gives only a flavour of the vast variety of reactions — whether acceptance, criticism, embellishment, emulation or emphatic rejection — that have proliferated in the primary literature. For example, he fails to dissect the complex blend of similarities and distinctions between *Tempo and Mode* and the much more recent model of punctuated equilibrium devised by Niles Eldredge and Stephen



Visionary: Simpson was an architect of a new approach to palaeontology.

Jay Gould, which posits rapid shifts during speciation events followed by prolonged stasis. Instead, Laporte relates the largely positive reaction to *Tempo and Mode*, as expressed mainly by ecologists and geneticists, in the years immediately after it was published. This truncated historical treatment creates the false impression that even some of Simpson’s most controversial proposals had rock-solid endurance. Modern approaches to the study of evolutionary rates are not simply a refinement of Simpson’s use of survivorship curves. To believe this is to ignore the substantial current debate over the assumptions made for these rate estimates and what they really explain. Laporte does offer a closing comment that some of the conclusions in *Tempo and Mode* are “badly flawed”, but he doesn’t tell us how.

Critical response to *Tempo and Mode* generally falls into two camps. Some applaud Simpson’s valiant steps into the *terra incognita* of evolutionary palaeontology, but substitute their own theories for his. Others remain dissatisfied with the empirical basis for some of either Simpson’s or more modern formulations. This resistance may partly stem from parochialism, but may also be a reasonable concern for what is justified by the evidence. As Laporte relates, even Dobzhansky and other biologists who heaped praise on *Tempo and Mode* when it first appeared remarked on a certain thinness of documentation.

In addition to describing the widespread interest in *Tempo and Mode*, the book shows how hugely influential and productive Simpson was in several other areas. Simpson the explorer is well documented, and Laporte skilfully links Simpson’s early quantitative studies of fossils that he himself collected with his later investigation of more general evolutionary problems. Simpson the bio-

geographer is aptly described as a rallying point for palaeontologists and geologists who rejected Alfred Wegener’s theories of continental drift. History, of course, shows this as the losing position. And the accumulation of evidence for plate tectonics in the 1960s eventually compelled Simpson to capitulate. As for his commitment to taxonomy and classification, Simpson spent much of his career describing many fossil mammals, and eventually produced a classification of the whole group.

Here again, Laporte unfortunately fails to place Simpson’s achievements in the context of more recent trends. While recognizing Simpson’s prodigious contributions, many today regard his mammalian classification as a step backward in concept and methodology. Moreover, Laporte hardly mentions Simpson’s conflicts with the rise of new approaches to systematic classifications, such as cladistics, and their threat to his philosophies.

In recounting Simpson’s personal and professional life, Laporte candidly offers a sad portrayal of a man who seemed to recognize with gratitude the warmth and support of his mentors but seemed unable to project those qualities himself. To be fair, Simpson’s intensive attention to his work and his enormous output would distinguish and, at the same time, distance anyone. Nonetheless, it seems that Simpson’s own reserve, whether a product of shyness or a sense of superiority, did not help to connect him closely with people.

The book relates the conflicts and resentment that induced Simpson to move to Harvard from the American Museum of Natural History. Laporte’s review of this matter, as well as his brief note on Simpson’s strained departure later from Harvard’s Museum of Comparative Zoology, lays much of the blame on Simpson himself. The biographer remarks that, despite the great man’s achievements, his “expectations were at times inflated, going beyond the bounds of what was reasonable in terms of other people’s needs and interests and responsibilities”.

Laporte derives some insights into Simpson’s mood in his final years from the brooding, melancholic Sam Magruder, the title character of Simpson’s posthumously published science-fiction novella. But it is clear that, in the Arizona retreat that served as home, library and laboratory in that final epoch, Simpson found much comfort in the companionship of his wife, Anne Roe, a statistician and clinical psychologist who, with Simpson, published the influential book *Quantitative Zoology* in 1939. He also had a small circle of friends and colleagues, and maintained his own indefatigable passion for scientific research until his death. ■

Michael J. Novacek is in the Division of Paleontology, American Museum of Natural History, New York, New York 10024, USA.

Biologists must take responsibility for the correct use of language in genetics.

We think we think with words: too often, the words think us. Four centuries ago, Francis Bacon showed us how those who say they seek knowledge are prisoners of language. He proposed the doctrine of the four idols — objects of false worship that block our understanding.

Bacon's terminology is unfamiliar, yet his formulations cut across recent fashions of thought and jolt us out of our preconceptions. The language we use about genetics and the genome project at times limits and distorts our own understanding, and the public understanding.

sequences previously unrecognized but so essential they have been conserved over hundreds of millions of years of evolution. Then, too, the entire phrase — the human-genome project: singular, definite, with a fixed endpoint, completed by 2000, packaged so it could be sold to legislative bodies, to the people, to venture capitalists. But we knew from the start the genome project would never be complete. The maps, or the sequences, are just the start of many lines of research, polyphiloprogenitive, multiply multiple genome projects.

The phrases current in genetics that most plainly do violence to understanding begin “the gene for”: the gene for breast cancer, the gene for hypercholesterolaemia, the gene for schizophrenia, the gene for homosexuality, and so on. We know of course that there are no single genes for such things. We need to revive and put into public use the term ‘allele’. Thus, “the gene for breast cancer” is rather the allele, the gene defect — one of several — that increases the odds that a woman will get breast cancer. “The gene for” does, of course, have a real meaning: the enzyme or control element that the unmutated gene, the wild-type allele, specifies. But often, as yet, we do not know what the normal gene is for.

Farbskizze eines narkotisierten *Drosophila melanogaster* Männchens

Außer dem linken Flügel ist alles normal

The sketch shows a male fruit fly from a side profile. The head is orange with large red eyes. The thorax is orange with black stripes. The abdomen is black with orange stripes. The left wing is deformed and labeled 'deformierter Flügel'. The right wing is normal and labeled 'Flügel'. Other labels include 'Fühler mit Arista' (antennae), 'Augen' (eyes), 'Schwingkölbchen' (halteres), 'Scutellum' (scutellum), and 'Tergit' (tergite).

Morgan and his group themselves understood what the language meant. At that time, a number of scientists were still intensely sceptical of mendelism. They did not believe in genes. In 1917, Morgan published an extended reply to these foes. This was just seven years after he had announced the discovery of the first mutant fruitfly — a male with white eyes instead of the normal red — and with it had demonstrated what we now call sex-linked inheritance.

“There are several reasons why we need the conception of the gene,” he wrote. In the first place, each gene has manifold effects. “If we take almost any mutant race” — or ‘strain’ as we would now say — “such as white eyes in *Drosophila*, we find that the white eye is only one of the characteristics that such a mutant race shows.” Further, Morgan wrote, although traits may vary, “there is at present abundant evidence to show that much of this variability is due to the external conditions that the embryo encounters during its development. Such differences as these are not transmitted in kind — they remain only so long as the environment that produces them remains.”

Pleiotropy. Polygeny. Perhaps these terms will not easily become common parlance; but the critical point never to omit is that genes act in concert with one another — collectively with the environment. Again, all this has long been understood by biologists, when they break free of habitual careless words. We will not abandon the reductionist mendelian programme for a hand-wringing holism: we cannot abandon the term gene and its allies. On the contrary, for ourselves, for the general public, what we require is to get more fully and precisely into the proper language of genetics. ■

CORNELLIA HESSE HONEGGER / PRO LITTERIS

The beanbag lives on

James F. Crow

The purpose of science, says Herbert Simon, "is to find meaningful simplicity in the midst of disorderly complexity". A real population includes individuals of all ages, some dying, some being born, some choosing mates, some reproducing, some migrating, some simply growing old. And in a sexual population each individual has a different genotype. Keeping track of all these goings-on is impractical and not very interesting. Most questions of genetic or evolutionary importance do not require such detail.

The gene-pool model is a wonderful, simplifying convention. The analogy is close to game-playing. A dictionary definition of a pool is "an aggregated stake to which each player has contributed". In the population model, each individual contributes to a large (theoretically infinite) pool of gametes. An offspring is a sample of two gametes from this pool. In the simplest case, the parents contribute equally to the pool and the sample is drawn randomly. For large diploid populations, this leads immediately to the familiar Hardy–Weinberg law (see Box below).

The simplicity and power of treating random sampling from a pool of genes as equivalent to random mating of diploid individuals is a bonus, especially appreciated by students in elementary genetics courses.

In a sexual population, each genotype is unique, never to recur. The life expectancy of a genotype is a single generation. In contrast, the population of genes endures. The quantities that are followed, in mathematical theories or in observations, are allele frequencies. The geneticist knows that at any



desired time, the genotype frequencies can be obtained by the simple binomial rule.

Mutation replaces one allele in the pool by another. Selection is introduced when parents contribute unequally to the gene pool. Meiotic drive entails preferential contribution of certain alleles in the gametes from a parent. Inbreeding and assortative mating (mating of phenotypically similar individuals) can be treated by correlations between the chosen gametes. Random sampling from a small parental population leads to random gene-frequency drift, because the same allele may be drawn twice. In the draw some alleles are over-represented, others under-represented, or not at all. These models lead to the simple mathematical equations of which classical population genetics largely consists.

Departures from simple assumptions can be met by convenient artifices. The most widely used is Sewall Wright's effective population number, N_e . In an ideal population each allele in the pool has an equal probability of coming from any parent. Departures from this, caused by uneven sex ratio or unequal fertility so that alleles do not have an equal probability of coming from any parent, are adjusted in the equations by using N_e — the size of an ideal population that would produce the same random drift or consanguinity as the actual population. This can be calculated from demographic studies of sex ratios, survival probabilities and fertility distributions, in addition to population size.

Ernst Mayr compared this type of gene-pool modelling to sampling from a bag of coloured beans and dismissed it as "beanbag genetics". Others have also criticized such models, especially those with a holistic bent, as being simplistic and gene-centred.

J. B. S. Haldane counterattacked with his spirited and witty paper "A defence of beanbag genetics", giving numerous examples, often from his own work, of the value of beanbag theory. Gene-pool models are indeed simplified, but this is "meaningful simplicity" that sweeps away "disorderly complexity". Often it is such a simplified view that provides the most useful insights into evolutionary processes. No one doubts that gene interactions in development are complicated. Fortunately, these processes are increasingly amenable to molecular techniques. At the same time the simplified models have allowed striking advances in the

Gene-pool

"In a sexual population, each genotype is unique, never to recur. The life expectancy of a genotype is a single generation. In contrast, the population of genes endures."

fields of evolution and population genetics, especially dynamic aspects.

Of course one can use more realistic equations. There may be larger units, such as linked clusters of genes, which, if not permanent, are enduring. This is especially important in molecular studies, where the unit of interest is a haplotype, as well as a nucleotide. Variable selection and geographically structured populations do not lend themselves to simple gene pool methods and have been modelled more realistically in other ways, of course with greater complexity.

In classical population genetics — as developed by R. A. Fisher, Haldane, Wright and, more recently, Motoo Kimura — the specific model was often left unspecified. These days, thanks to the pioneering work of Gustave Malécot, models are stated much more precisely. The idea is to specify the model in enough detail for the equations, whether deterministic or stochastic, to follow unambiguously. A mathematical model is based on precisely stated assumptions as to population size, ploidy, extent of self-fertilization, rules of migration, time and nature of mutation, patterns of selection and mating combinations. Investigators using an equation then know that for at least one specified model the equation is correct, and they can judge how well this model mimics nature. In modern research, equations are very sophisticated, computers being available when the mathematics become prohibitive. As a means of bypassing distracting details to provide useful approximations and novel insights, gene-pool models are here to stay. ■

James F. Crow is in the Genetics Department, University of Wisconsin, Madison, Wisconsin 53706, USA.

FURTHER READING

Dawkins, R. *The Selfish Gene* Revised edn (Oxford Univ. Press, 1989).
Fisher, R. A. *The Genetical Theory of Natural Selection* Complete variorum edn (Oxford Univ. Press, 1999).
Haldane, J. B. S. A defence of beanbag genetics. *Persp. Biol. Med.* **7**, 343–359 (1964). Reprinted in *Selected Genetic Papers of J. B. S. Haldane* (ed. Dronamraju, C. R.) 1–17 (Garland, New York, 1990).
Nagylaki, T. Gustave Malécot and the transition from classical to modern population genetics. *Genetics* **122**, 253–268 (1989). Reprinted in *Perspectives on Genetics* (eds Crow, J. F. & Dove, W. F.) 103–118 (Univ. Wisconsin Press, Madison, 2000).

The Hardy–Weinberg law

The Hardy–Weinberg (HW) law relates the frequencies of genes to those of genotypes in a randomly mating population. It is an application of the binomial expansion $(p + q)^2 = p^2 + 2pq + q^2$, where p and q are the proportions of alleles A and a , and p^2 , $2pq$ and q^2 are the proportions of genotypes AA , Aa and aa . If the parents mate at random, which is the equivalent to combining genes at random from a large pool to which each parent has contributed equally, the zygotes are in HW proportions. The larger the population, the closer the numbers agree with these binomial expectations. With differential mortality among the genotypes, the population will no longer be in HW proportions. In fact, departures from HW proportions are often taken as evidence for selection. With more than two alleles, the extension is straightforward; the binomial expansion becomes multinomial.

A material fix

Richard P. Wool

The idea of materials that can mend themselves seems far-fetched. But a system that allows composite materials to 'self-heal' has been devised and has passed some early tests.

Everyone has had minor scratches or abrasions that have healed, the damaged skin becoming as good as new. In contrast to that, we are now so used to products of all sorts breaking or wearing out that we fully accept planned obsolescence, disposable or busted materials, and mountains of rubbish looking for a landfill site. Materials are bound to become fatigued and weakened, however, and self-healing in which the material is made as good as new would seem to be the stuff of fantasy. Not so, say White and co-workers on page 794 of this issue¹, in a paper entitled "Autonomic healing of polymer composites".

Polymer composites are advanced materials that consist of two components — a reinforcing fibre (such as carbon, glass, Kevlar, flax, hemp or jute) and a liquid moulding resin (epoxy, unsaturated polyester, vinyl ester or urethane). Typically, the fibres are first placed in a mould. The resin, which has a viscosity similar to that of olive oil, is poured or injected into the mould. The resin hardens, and out comes the finished item. We use these composite materials in all manner of products, for all manner of applications. The market is huge: some 20 million tonnes of composites are used annually in civil engineering, aerospace and defence-related projects, offshore oil exploration, electronics and biomedicine.

So, how are materials repaired — and why do they fail in the first place? Biological healing processes are exceedingly complex and involve events that occur both simultaneously and sequentially, and are affected by a host of external variables (temperature, humidity and pressure, as well as electrical, ultrasonic and magnetic fields). Healing of materials is similarly complex and involves several stages which together act to eliminate damage induced by mechanical, chemical or thermal means. The damage may involve bond rupture, the formation of microvoids, debonding of the fibres from the resin, or various other events at the molecular and microscopic levels that collectively weaken the material and perhaps put its user at risk. The most common route to failure is fatigue — the formation of microvoids incurred by mechanical stress through repeated usage. The microvoids enlarge and coalesce to form microcracks in the material; in turn the microcracks can grow to catastrophic proportions and cause total failure of the product.

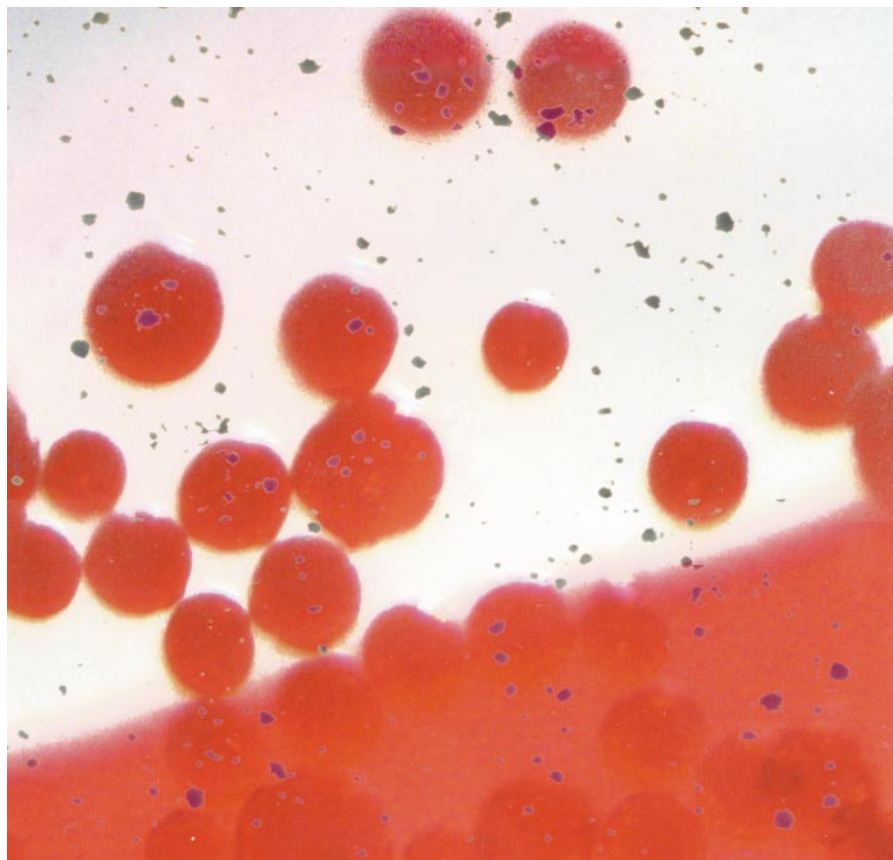


Figure 1 Autonomic healing¹ in action. This image shows a crack in an epoxy polymer material, the crack plane running along the bottom of the picture. The red spheres are microcapsules, 200 micrometres in diameter, containing the red-dyed repair fluid. Following polymerization by Grubbs' catalyst, the fluid has filled part of the crack. The material being mended appears white. The black flakes are the Grubbs' catalyst, embedded in the material's matrix. (Picture courtesy of S. R. White *et al.*)

As I have described elsewhere², various workers have studied crack repair in polymers using thermal or solvent processes. In these processes, the polymer is softened by heating or with a solvent, and the cracks weld themselves in several stages. But applying the same treatment to a large composite structure is not very practical.

The challenge for White *et al.*¹ was to design a self-repairing composite to satisfy the following wish list. It has to contain a self-repair system that does not affect the material's overall properties or performance. It must also be able to sense damage, and then be able to react to that damage and initiate healing. Finally, it must restore the material's original properties (strength and stiffness, for example).

This is a formidable array of require-

ments, which is probably why White *et al.* are the first to claim success in meeting them. Their system is ingenious and involves adding several low-volume components to the composite. It is outlined diagrammatically in their Fig. 1 on page 794.

Microcapsules are first filled with a repair fluid (dicyclopentadiene, or DCPD). This fluid can polymerize in the presence of a catalyst (patented by R. Grubbs, and known as the Grubbs' catalyst) to form a very tough polymer. The catalyst and the fluid must not meet until they are needed in the damaged zone, which is the function of the microcapsule membrane. The catalyst is distributed first in the epoxy resin, followed by the microcapsules filled with the DCPD fluid. This modified resin is added to the fibres,

the epoxy hardens, and the composite is moulded into its final shape.

Testing the system involved mechanically loading the composite, causing microcracks to form in the matrix of the epoxy polymer. What is especially clever is that, for stress field reasons, the microcracks migrate towards the soft, fluid-filled microcapsules and rupture them, releasing the repair fluid. The fluid travels down the microcracks by capillary action and encounters the Grubbs' catalyst, which initiates the polymerization reaction. The polymerized fluid fills the microcracks and repairs them (Fig. 1), largely restoring the composite material to its original strength and stiffness.

The concept of self-healing composites cuts against the grain of current material-design principles, but has far-reaching con-

sequences for improving product safety and reliability. The approach should prove especially useful where it is not possible, or practical, to repair the material once it has been put into use. Components of vehicles used in deep-space exploration, satellites, rocket motors and prosthetic organs are prime candidates for such treatment. So too are the space stations of the present and future, and engineering innovations such as bridges constructed from composite materials. ■

Richard P. Wool is at Cara Plastics Inc., and is in the Affordable Composites from Renewable Resources Program, Center for Composite Materials, University of Delaware, Newark, Delaware 19716-3144, USA. e-mail: wool@ccm.udel.edu

1. White, S. R. *et al.* *Nature* **409**, 794–797 (2001).

2. Wool, R. P. *Polymer Interfaces: Structure and Strength* (Hanser/Gardner, Munich, 1995).

and Rosen to suggest that quantum mechanics is incomplete, on the basis that any theory of nature must be both 'local' and 'realistic'. Local realism is the idea that, because the properties of one particle cannot be affected by a particle that is sufficiently far away, all properties of each particle must exist before they are measured. But non-separability contradicts this sort of local realism.

For a long time, this debate about the nature of quantum mechanics was largely theoretical, but in 1964 Bell introduced a set of mathematical equations — Bell's inequalities — that could be tested experimentally. The inequalities must be obeyed by any local realistic (classical) theory, whereas they are violated by quantum mechanics. It became clear at the beginning of the 1980s that the overwhelming majority of experiments supported the predictions of quantum mechanics. But two important experimental loopholes meant the evidence was inconclusive — at least to supporters of local realism.

The first of these loopholes, the so-called locality loophole, arises whenever measurements are performed on two spatially separated particles, because any possibility of communication between the two parts of the apparatus must be excluded. This can be achieved if the two measurements occur in different 'light cones', which means they cannot be connected by a signal travelling

Quantum physics

Count them all

Philippe Grangier

Direct experimental evidence to resolve the conflict between classical and quantum physics has been a long-awaited goal. As the last loophole closes, it seems that quantum mechanics was right all along.

Certain features of quantum mechanics are undeniably strange, and it has taken a long time for physicists to understand them — a necessary step before being able to accept them. The predictions of quantum physics often seem to contradict the intuitive, and seemingly reasonable, assumptions about how the world should behave that can be deduced from classical physics. Not surprisingly, many experiments have been done to explore these differences, and the results have tended to support quantum mechanics. But there have been persistent loopholes in these experiments, allowing alternative interpretations of the data. On page 791 of this issue, however, Rowe *et al.*¹ describe how they measured the quantum correlations between two beryllium ions with nearly perfect detection efficiency — thereby closing the last experimental loophole, and providing compelling evidence that some basic ideas inherited from classical physics must be abandoned.

Non-separability, or entanglement, has emerged as the most emblematic feature of quantum mechanics. Briefly stated, it says that one can define 'entangled' quantum states of two particles in such a way that their global state is perfectly defined, whereas the states of the separate particles remain totally undefined. In other words, the information contained in such an entangled state is all about the correlation between the two particles, and nothing is said — even more, nothing can be known — about the states of the individual particles.

Some thinking is necessary to realize how weird this is. In 1935 it led Einstein, Podolsky

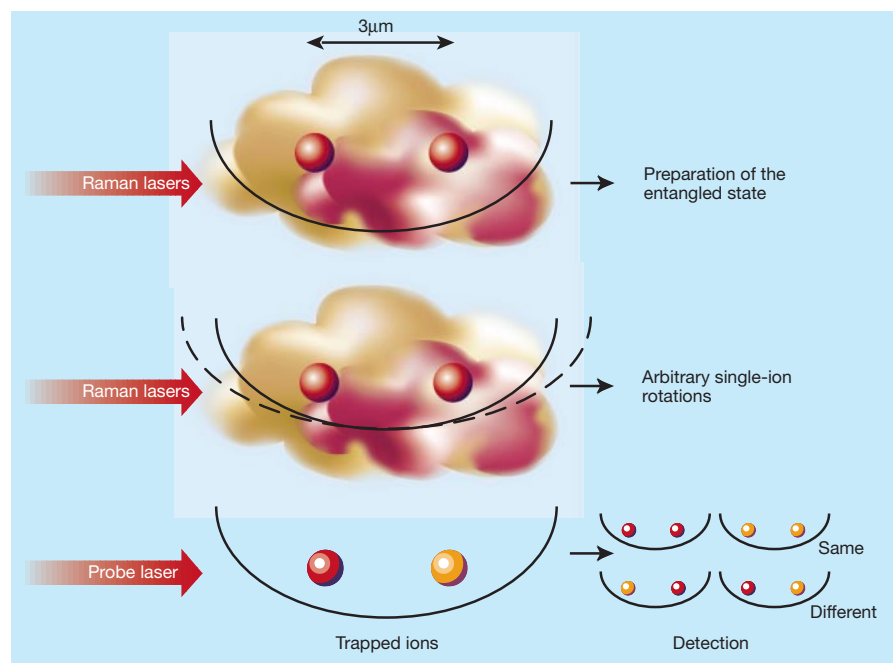


Figure 1 Testing our quantum picture of the world. Rowe *et al.*¹ prepared two ions in an entangled state by using Raman laser beams. Next, they applied an arbitrary rotation to each ion by simultaneously controlling the phase of the Raman beams and the strength of the ion trap. Finally, they determined the quantum state of each ion by shining light (the probe laser) on the ion pair to identify whether each ion fluoresces. There are four possible outcomes: two in which the states are the same (both ions are bright or both are dark), and two in which they are different (only one ion or the other fluoresces). The measurements give the ratio of $(N_{\text{same}} - N_{\text{different}}) / (N_{\text{same}} + N_{\text{different}})$, where N is the number of detected ion pairs. The experimental result violates Bell's inequalities — thereby confirming the predictions of quantum mechanics — in a way that escapes the 'detection-efficiency' loophole.

at a speed equal to or less than the velocity of light. There were considerable difficulties with doing this sort of experiment, but finally all were solved, and quantum mechanics, once again, turned out to be right².

There still remained a second loophole, based on the statistical nature of previous experiments. The 'detection-efficiency' loophole argues that, in most experiments, only a very small fraction of the particles generated are actually detected. So it is possible that, for each measurement, the statistical sample provided by the detected pairs is biased. For example, in experiments using pairs of photons emitted by atomic cascades², only one pair in every million was used in the measurement. So, to extract a meaningful conclusion from the observed data, it was necessary to assume that a small fraction of data provides a fair statistical sample. This would be similar to allowing one hundred votes to decide a ballot of one hundred million. Not surprisingly, several local realistic models were built to mimic the experimental results, and the detection-efficiency loophole became the Achilles' heel of experimental tests of Bell's inequalities.

It was first thought that improving the detection efficiency in experiments with pairs of entangled photons² would solve this problem. But this proved more difficult than expected and, despite several proposals^{3,4}, no conclusive experiment has been done. Another avenue is to use correlated measurements on pairs of massive particles, with the hope that their quantum states would be easier to detect than those of photons. Theoretical proposals considered pairs of atoms produced through a photodissociation process^{5,6}, or pairs of Rydberg atoms⁷. But such experiments are not easy, and no conclusive test of Bell's inequalities has been carried out in these systems.

In their experiment, Rowe *et al.*¹ manipulate the quantum states of two massive entangled particles — in this case trapped ions. For an efficient measurement, it is necessary to detect and compare the individual quantum states of the two ions, which are separated by only a few micrometres. The authors use a two-step measurement process, which results in four possible outcomes (Fig. 1). Contrary to the photon-correlation experiments, in which many photon pairs are missed, here every ion pair is included in the measurement. Remaining errors are attributed to an incorrect preparation of the initial state (less than 12%), a phase error in the rotations (less than 6%), or a mistake in the identification of the final states (less than 2%). The agreement with quantum mechanics is excellent, making this the first violation of Bell's inequalities with high enough efficiency (80% overall) to close the detection loophole.

What conclusions can be drawn from this experiment? It does not close the locality

loophole, whereas the experiments that do, cited in ref. 2, do not close the efficiency loophole. Closing both loopholes in the same experiment remains a challenge for the future, and would lead to a full, logically consistent rejection of any local realistic hypothesis. Even so, the overall agreement with quantum mechanics seen in all the experimental tests of Bell's inequalities is already outstanding. Moreover, each time a parameter is changed that was considered to be crucial (for example, using time-varying measurements, or increasing the detection efficiency), the experiments show that these changes have no consequence: the results continue to agree with quantum-mechanical predictions. This appears rather compelling evidence to me that quantum mechanics is right, and cannot be reconciled with classical physics.

Finally, Rowe *et al.*'s experiment is a vivid illustration of the high degree of sophistication that has been reached in the

control of 'engineered' quantum systems. The entangled-ion pair was prepared on purpose, and can be observed at will. This opens fascinating possibilities for manipulating the quantum state of entangled many-particle quantum systems. Such systems are the basic elements for achieving long-term goals in quantum information processing, such as building a quantum computer. ■

Philippe Grangier is at the Laboratoire Charles Fabry, Institut d'Optique Théorique et Appliquée, F-91403 Orsay, France.

e-mail: philippe.grangier@iota.u-psud.fr

1. Rowe, M. A. *et al.* *Nature* **409**, 791–794 (2001).
2. Aspect, A. *Nature* **398**, 189–190 (1999).
3. Kwiat, P. G., Eberhard, P. H., Steinberg, A. M. & Chiao, R. Y. *Phys. Rev. A* **49**, 3209–3220 (1994).
4. Huelga, S. F., Ferrero, M. & Santos, E. *Phys. Rev. A* **51**, 5008–5011 (1995).
5. Lo, T. K. & Shimony, A. *Phys. Rev. A* **23**, 3003–3012 (1981).
6. Fry, E. S., Walther, T. & Li, S. *Phys. Rev. A* **52**, 4381–4395 (1995).
7. Freyberger, M., Aravind, P. K., Horne, M. A. & Shimony, A. *Phys. Rev. A* **53**, 1232–1244 (1996).

Ecology

The rising cost of bushmeat

Peter D. Moore

Some plants depend on specific animal vectors for the dispersal of their seeds. If the vector comes under threat, there are likely to be adverse consequences for the plant.

As human pressure on tropical forests increases, so does the extent of bushmeat hunting — the killing of wild species for food. The adverse effects go beyond the species immediately concerned, and a striking example can be found in a paper by L. F. Pacheco and J. A. Simonetti in *Conservation Biology* (14, 1766–1775; 2000).

As might be expected, large-bodied animals — including some primates — are usually the preferred bushmeat prey. Many of these animals have fruits and seeds as a major part of their diet; in turn, the dispersal of the plants concerned may depend on the passage of seeds through the animals' guts, and subsequent deposition elsewhere. Most fruit-eaters are generalists, so if one vector is lost there may be others to take its place. But in fruit consumption, size matters, and the selective loss of all large frugivores could seriously impair dispersal in plants with large fruits. Moreover, some plants are relatively specific in their seed-dispersal vectors, and so are most at risk.

Pacheco and Simonetti have studied one such species, *Inga ingoides* (family Fabaceae), which is one of the common trees in the lowland forests of Bolivia. The seeds

Figure 1 Hunted look: the spider monkey of South America.



TOM BRAKEFIELD/CORBIS

of this tree are dispersed almost exclusively by the spider monkey (*Ateles paniscus*; Fig. 1). Three other primate species feed upon the tree's seeds in the study area, but only spider monkeys ingest them intact and act as effective dispersal agents. The monkeys can travel 1 kilometre in a few minutes and have a gut passage time of about 4 hours. So the seeds can be dispersed some distance from the parent plant, ensuring that the genetic structure of the population of *I. ingoides* is well mixed.

The research concentrated on the genetic variability of *I. ingoides* populations, comparing areas where spider monkeys were present with those where they had become locally extinct through bushmeat hunting. After analysing 14 enzyme systems in leaf extracts, the authors found that there was less genetic variation in the seedling populations around parent trees when the monkeys were absent than when they were present. Without the monkeys the seeds simply fall to

the ground around their parents. In other words, the monkeys are responsible for maintaining a thorough genetic mix in the population, and in their absence a series of genetically more uniform patches develops.

In general, trees in the tropical rainforest have high levels of genetic diversity. They are usually out-breeding and have efficient gene flow because of their specialized pollination and seed-dispersal mechanisms. Fragmentation of the forest, whether by physical processes (such as clearance) or by interruption to gene flow (as in the case of vector loss), can lead to the local accumulation of potentially detrimental mutations and an overall loss of fitness. As has so often proved to be the case in ecological studies, when one link is removed from a network, the whole system can start to unravel. ■

Peter D. Moore is in the Division of Life Sciences, King's College, Franklin-Wilkins Building, 150 Stamford Street, London SE1 9NN, UK.
e-mail: peter.moore@kcl.ac.uk

Biochemistry

Single-handed cooperation

Jay S. Siegel

Our bodies use only 'left-handed' amino acids and 'right-handed' sugars. Hints are now emerging on how this handedness evolved and how cooperativity among like-handed molecular components came about.

The sequencing of the human genome provides exciting possibilities to explain the complexities of life. But some more basic questions remain unanswered — such as why the double-helix structure of DNA spirals in a clockwise (right-handed) direction, rather than a left-handed one. On page 797 of this issue¹, Ghadiri *et al.* use a peptide system to demonstrate how 'homochirality', or single-handedness, may have evolved in biological molecules.

Nucleic acids, like nearly all biological molecules, exhibit 'handedness', and only one of the two forms is used in a particular biological process. This gives rise to homochirality, where each molecule is identical. Among sugars, ribose and deoxyribose are not identical; nonetheless, the form of ribose in RNA and the form of deoxyribose in DNA share a common right-handed orientation of their principal functional groups. This common stereochemistry — the three-dimensional shape of the molecule — evokes another concept of homochirality, which relates members of a family of compounds.

The natural amino acids share a common stereochemistry, as they are all left-handed (L-amino acids). This raises the question of whether homochirality within a family of biological molecules is the result of a stereochemical cooperativity (diastereo-

selectivity) among the members within a biopolymer. Indeed, our enzymes and nucleic acids are composed of predominantly L-amino acids and D-sugars — we are unable to use the opposite-handed biopolymers. Why not? Attempts to answer this question by mimicking the creation of a homochiral environment in the laboratory have been unsuccessful until now.

This difference in functional chiral form had tragic consequences in the 1960s when pregnant women were given a sedative (Thalidomide) that was a mixture of the right- and left-handed forms of the drug; one of the two forms, or enantiomers, gave rise to birth defects. That two enantiomers can have such different functions shows that stereochemistry controls whether, and how, molecules recognize one another and 'shake hands'.

The origin of one-handedness in biological molecules is not yet clear². Several explanations have been put forward to explain how homochirality came about, but all are speculative — it is not even known yet whether it arose by chance or by some other means^{3,4}.

Synthetic polymers are simpler than biological systems and provide a model for understanding the origin of homochirality in biomolecules. One proposal stems from observations that polymers made from



100 YEARS AGO

Sensational Newspaper Reports as to Physiological Action of Common Salt.

In the interest of the dignity of scientific research I venture to hope you will print the following statement. Some American papers have recently published sensational and absurd reports of physiological theories and experiments whose authorship they attributed to me. These reports, which in America nobody takes seriously, were reprinted and discussed in European papers. I hardly need to state that I am in no way responsible for the journalistic idiosyncrasies of newspaper reporters and that for the publication of my experiments or views I choose scientific journals and not the daily Press.

Jacques Loeb

From *Nature* 14 February 1901.

50 YEARS AGO

A recent address on "Freedom in Science" broadcast by Prof. C. A. Coulson made even clearer the need for ensuring that our instruments of government make proper provision for such freedom. Prof. Coulson observed that, since patience, humility, tolerance, fairmindedness, integrity, co-operation and trust are the hallmarks of the scientific tradition, he has been forced to the conclusion that this tradition is ultimately based on, and derives its final sanction from, moral and ethical considerations which lie outside the field of what is popularly called science. A. N. Whitehead maintained that the inner conviction that the world is rational, and the confidence in one another that enables us to dispense with the verification of other people's claims, is a legacy coming to us from "the medieval insistence on the rationality of God". Prof. Coulson accordingly suggests that in a time of crisis like the present, which is marked by the breakdown of personal conviction among ordinary people, it is highly important that... men of science are the custodians of many of the most precious values of our civilization. From this aspect, Prof. Coulson agrees with Prof. Polanyi's view that the suppositions underlying our belief in science "co-extend with the entire spiritual foundations of man, and go to the very root of his social existence". But he discusses a possible danger to the freedom of science which might well arise from intoxication with power and an unmeasured faith in organization. This will have to be guarded against when the Science Centre in London now under consideration is established.

From *Nature* 17 February 1951.

building blocks of random handedness will contain mixtures of right- and left-handed blocks so complex that no two polymers will have identical stereochemistry. All the polymers will be chiral, but if one exists, it is unlikely that its mirror image will too⁵. For small polymers consisting of only a few building blocks, the number of possible combinations of right- and left-handed blocks is small, and the mirror images are easily formed. However, for a polymer comprising 20 building blocks there are almost a million possibilities, and an enormous number of blocks would be required to build all the possible mirror images. Biological molecules often have over 100 building blocks, pushing the limits of available materials and making it extremely unlikely that a molecule and its mirror image can be prepared in the same batch. If the sample of polymers contained some that were self-replicating, it is reasonable that the most efficient one will emerge, and only this homochiral polymer will exist⁶.

In complex organisms and living systems,

homochirality manifests itself in families of related biological molecules, such as the naturally occurring L-amino acids in a protein. The presumption here is that there is cooperativity among these biomolecules in their biopolymers or underlying structure. The polymer example above alludes to an optimum self-replicating polymer. Can we form a general rule that homochirality leads to optimal function? With regard to self-replication, this question can be addressed by stereochemical analysis and reaction engineering. However, engineering the parameters of biochemical processes to obtain appropriate stereoselectivity — choosing of the preferred mirror image — has not been trivial. Poor reaction selectivity and inhibition of replication have plagued researchers who would like to mimic life through protobiology⁷.

Ghadiri *et al.*¹ find a stereoselective reaction, between two chiral peptides, that uses the product of their coupling to replicate — a process known as 'autocatalysis'. Their results indicate that this self-replicating

Daedalus

David Jones

David Jones, author of the Daedalus column, is indisposed.

system could perpetuate homochirality, because a left-handed template is competent to bring together only those fragments that are also left-handed. The authors also show that, even if only one out of 15 building blocks has opposite handedness, autocatalysis is significantly diminished. Such a result establishes that homochiral peptides are most efficient at autocatalysis.

The authors also found that even though a single 'mutation' with regard to handedness could hamper autocatalysis, the same mutant template was reasonably effective at catalysing the combination of two pure left-handed fragments. The combination of these results supports the idea that a polymer evolution experiment would ultimately result in homochirality of biological molecules. Furthermore, these results support a general principle that similarity in building blocks leads to higher-order structure, like the spiral shape of a ram's horn or a crystal-lattice structure^{8,9}.

Given the fundamental nature and pervasive occurrence of chirality in biological and non-biological macromolecules, it is reasonable to postulate that homochirality existed long before the genetic information of life was encoded. Once in place, this sense of chirality perpetuated throughout early life systems and has become a part of complex biochemical processes such as translation and transcription. So the linear sequence of information on the genome is passed along owing to the three-dimensional structure of the DNA double helix, and must now include the coding for homochirality that we see in amino acids. We now have a model system that may bring us closer to understanding why we see cooperativity among homochiral biological molecules today.

Jay S. Siegel is in the Department of Chemistry, University of California San Diego, La Jolla, California 92093-0358, USA.
e-mail: jssiegel@ucsd.edu

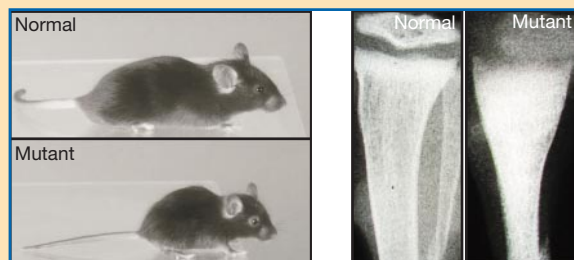
1. Saghatelian, A., Yokobayashi, Y., Soltani, K. & Ghadiri, M. R. *Nature* **409**, 797–801 (2001).
2. Bonner, W. A. *Orig. Life Evol. Biosph.* **25**, 175–190 (1995).
3. Siegel, J. S. *Chirality* **10**, 24–27 (1998).
4. Berger, R., Quack, M. & Tschumper, G. S. *Helv. Chim. Acta* **83**, 1919–1950 (2000).
5. Green, M. M. & Garetz, B. A. *Tetrahedr. Lett.* **25**, 2831–2834 (1984).
6. Bolli, M., Micura, R. & Eschenmoser, A. *Chem. Biol.* **4**, 309–320 (1997).
7. Kozlov, I. A., Politis, P. K., Pitsch, S., Herdewijn, P. & Orgel, L. E. *J. Am. Chem. Soc.* **121**, 1108–1109 (1999).
8. Kuhn, H. & Waser, J. *Angew. Chem. Int. Ed. Engl.* **20**, 500–520 (1981).
9. Thompson, D. *On Growth and Form* 346 (Cambridge Univ. Press, 1961).

Medicine

Channel fault in osteopetrosis

Bone is a dynamic structure, constantly being formed (by cells called osteoblasts) and resorbed (by osteoclasts). The activities of these cells must be finely balanced. If resorption exceeds formation, the result is the weakened, brittle bones characteristic of osteoporosis. In osteopetrosis, by contrast, not enough bone is resorbed: the osteoclasts malfunction, leading to dense yet fragile bones with no bone marrow. Writing in *Cell* (**104**, 205–215; 2001), Uwe Kornak, Dagmar Kasper and colleagues identify a molecule that is needed for the bone-resorbing cells to function. Their results may explain some cases of an inherited disease, infantile malignant osteopetrosis.

During bone resorption, osteoclasts attach to the bone matrix and form a specialized outer membrane, called the ruffled border, facing the bone. Bone-digesting enzymes are then transported out of the cell, across this border. The enzymes need an acidic environment to function, so the osteoclasts also pump out protons. At the same time, chloride ions are let out of



the cell to maintain the electrical balance. The protein-based channel that selectively allows the efflux of chloride ions has been elusive, but the authors suggest that it is a molecule called CIC-7.

Kornak *et al.* engineered mice with a disruption in the gene encoding CIC-7. The resulting mutant mice (one of which is pictured above, with a normal mouse) had all the characteristics of osteopetrosis, including short limbs and abnormally dense bones. The mice also failed to form bone marrow, as can be seen by comparing the two high-magnification X-ray images above. Moreover, osteoclasts isolated from the mutant mice were unable to absorb bone.

The authors tracked down the CIC-7 protein to the outer membrane and certain acidic organelles, called lysosomes, in the osteoclasts of normal mice. But there was no CIC-7 protein in these cells in the mutant mice. Although the osteoclasts from the mutant mice contained the usual proton pump in their outer membranes, they were unable to acidify the extracellular space — presumably because of the defect in CIC-7. Finally, Kornak *et al.* discovered that a patient with infantile malignant osteopetrosis has a disrupted CIC-7 gene. All in all, they make a convincing case that this chloride channel is crucial for bone resorption.

Amanda Tromans

Obituary

Konrad E. Bloch (1912–2000)

Konrad Bloch, who died on 15 October 2000, was one of the small group of biochemists who first used stable isotopes to study the biological synthesis of complex molecules. His brilliant dissection of how cholesterol is made in living tissues was a remarkable achievement, especially given the state of knowledge in the 1940s and 1950s of how living organisms synthesize molecules.

Bloch was born in 1912 in Neisse, eastern Germany (now Nysa in Poland). When the Nazis came to power in 1933, he was a chemistry student in Hans Fischer's laboratory at the Technische Hochschule in Munich. Bloch, who was Jewish, was told that he was ineligible to continue at the institute because Fischer had declined to accept him as a graduate student. But this was a lie. It was the new racial laws that blocked further study and, having gained his degree in chemical engineering, Bloch left Germany in 1934.

After a brief stay in Switzerland, and with the help of R. J. Anderson at Yale, Bloch moved to Columbia University, New York, where he received his PhD in biochemistry. He joined the laboratory of Rudolph Schoenheimer, another refugee, who had started to use stable isotopes to trace biochemical pathways. After Schoenheimer's untimely death in 1941, his students divided the work among themselves and, by chance, Bloch inherited lipids as his research area.

Bloch made rapid progress in his studies on the biogenesis of cholesterol. With David Rittenberg at Columbia, then in his own laboratory at Chicago, and from 1954 at Harvard, Bloch explored the synthetic pathway that leads from acetic acid to cholesterol and involves more than 30 reactions. A major breakthrough was the prediction that a symmetrical C_{30} hydrocarbon, squalene, was an intermediate in the pathway. Bloch's group subsequently showed that squalene could be cyclized to form the sterol called lanosterol. In 1953, Bloch and the organic chemist R. B. Woodward proposed a mechanism for this cyclization reaction that turned out to be largely correct.

But squalene formation needs an isoprenoid precursor, and the identity of this elusive molecule was unexpectedly revealed at the Merck Sharp & Dohme laboratories during studies on bacteria that do not synthesize sterols. The six-carbon mevalonic acid was quickly recognized as the possible missing link in sterol biosynthesis, because eliminating



Research chemist who outlined the path of cholesterol synthesis

a carboxyl group and water from the molecule generated the necessary isoprenoid precursor for squalene.

Bloch's laboratory made some of the key findings, with important contributions also coming from the joint efforts of John Cornforth and George Popják in Britain, and from Feodor Lynen's group in Germany. Eventually, two phosphorylated compounds arising from mevalonic acid were identified as the biological isoprene units. Six of these five-carbon isoprenes are joined together to form squalene, which is then converted to the 27-carbon cholesterol. Bloch's work in this area was recognized by the award of the Nobel prize in Physiology or Medicine, which he shared with Lynen in 1964. Cornforth received a Nobel prize in Chemistry in 1975.

Information about the pathway from acetic acid to cholesterol greatly aided the later discovery of statins, drugs that interfere with cholesterol synthesis. These drugs are now widely used to treat high levels of cholesterol in the blood.

Bloch was a scholar who was driven by curiosity. He maintained his interest in cholesterol throughout his life, and in the 1970s and 1980s provided an explanation for why cholesterol, rather than other sterols, had evolved as a functional component of cellular membranes.

But as a young scientist Bloch also made substantial contributions to the area of peptide synthesis. He showed that in yeast, turning a saturated fatty acid (one which

contains no carbon–carbon double bonds) into an unsaturated fatty acid requires molecular oxygen. This led him to wonder how bacteria that grow only in the absence of oxygen are able to make unsaturated fatty acids. His studies led him to propose an anaerobic pathway for the formation of the carbon–carbon double bond. Extension of this work led to the discovery of an acetylenic 'suicide substrate', one of the first of this class of enzyme inhibitors. Bloch once stated that his only regret in science was that he and many co-workers had wasted so much time in the intractable process of trying to purify from membranes the enzymes involved in fatty acid and cholesterol biosynthesis.

Bloch was a dedicated teacher. For decades he taught the basic biochemistry course at Harvard. Thousands of students, many of whom are now scientists, physicians and scholars themselves, first learned their biochemistry from him. Bloch trained over 125 graduate and postdoctoral students, virtually all of whom remember the time in his laboratory warmly and as being formative in their careers. His gentle, subtle humour, usually accompanied by a distinct twinkle in his eyes, was keen but rarely destructive. His habit was to make the rounds of his co-workers at least once a day, asking what their latest findings were. Any promising result was discussed immediately and the implications for further experiments were explored. But he was not a hard taskmaster — the spirit in his laboratory was one of cooperation rather than competition.

As a result of a cultured middle-class upbringing in Germany between the two world wars, Konrad Bloch had a lifelong interest in music and the visual arts. His fascination in chemistry blossomed in Fischer's laboratory after finding the study of metallurgy uninspiring. He was then led to biochemistry through his early contacts with lipid research in Switzerland, but most importantly through his association with Schoenheimer. He is survived by Lore to whom he was married for 59 years, a son Peter, a daughter Susan, and two grandchildren.

Howard Goldfine and Dennis E. Vance

Howard Goldfine is in the Department of Microbiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104-6076, USA.

e-mail: goldfinh@mail.med.upenn.edu

Dennis E. Vance is in the CIHR Group on Molecular and Cell Biology of Lipids, and the Department of Biochemistry, 328 Heritage Medical Research Centre, University of Alberta, Edmonton, Alberta T6G 2S2, Canada.

e-mail: dvance@pop.srv.ualberta.ca

http://www.nobel.se/medicine/laureates/1964

Growth of domesticated transgenic fish

A growth-hormone transgene boosts the size of wild but not domesticated trout.

Growth rates of many fish species used in aquaculture are naturally slow, but are currently being enhanced by traditional methods of domestication and selection¹. The efficiency of growth and feed-conversion can also be increased in finfish by creating transgenic fish that incorporate a gene construct encoding growth hormone, giving 3–11-fold gains in weight^{2–5}. Here we examine growth enhancement due to transgenesis in wild (slow-growing) and selected (fast-growing) commercial salmonid species relative to growth achieved in domesticated strains. We find that the growth response is strongly influenced by the intrinsic growth rate and genetic background of the host strain, and that inserting growth-hormone transgenes into highly domesticated fish does not necessarily result in further growth enhancement.

We microinjected rainbow-trout (*Oncorhynchus mykiss*) eggs from a very slow-growing wild strain with a salmon gene construct overexpressing growth hormone (construct *OnMTGH1*). Like coho salmon³, the transgenic trout grew much faster than non-transgenic sibling controls (Fig. 1a), achieving a 17.3-fold difference in weight by 14 months post-fertilization (non-transgenic: fish weight, 9.7 ± 0.6 g; length, 9.3 ± 0.1 cm; $n=247$; transgenic fish: weight, 167.6 ± 14.9 g; length, 24.2 ± 0.9 cm; $n=40$). The amount of growth enhancement was comparable (17.6-fold) in a first-generation (F_1) family line established from these fish.

However, the growth of transgenic wild-strain rainbow trout did not surpass that of a fast-growing non-transgenic domesticated strain of trout used in aquaculture (Fig. 1a). We found that introducing the growth-hormone construct into this domestic strain (Fig. 1a) did not cause further growth enhancement (non-transgenic fish: weight, 269.4 ± 5.9 g; length, 24.0 ± 0.22 cm; $n=151$; transgenic fish: weight, 282.7 ± 26.2 g; length, 25.8 ± 1.58 cm; $n=9$). These results indicate that similar alterations of growth rate can be achieved both by selection and by transgenesis in rainbow trout, but that the effects are not always additive.

Transgenic wild-strain rainbow trout retain the slender-body morphology of the wild-type strain (Fig. 1a), but their final size at sexual maturity can be affected by transgenesis (Fig. 1b), an observation that warrants concern⁶. Domestic and wild-strain transgenic trout had reduced viability, and, in the case of the domestic strain, all transgenic animals died before sexual

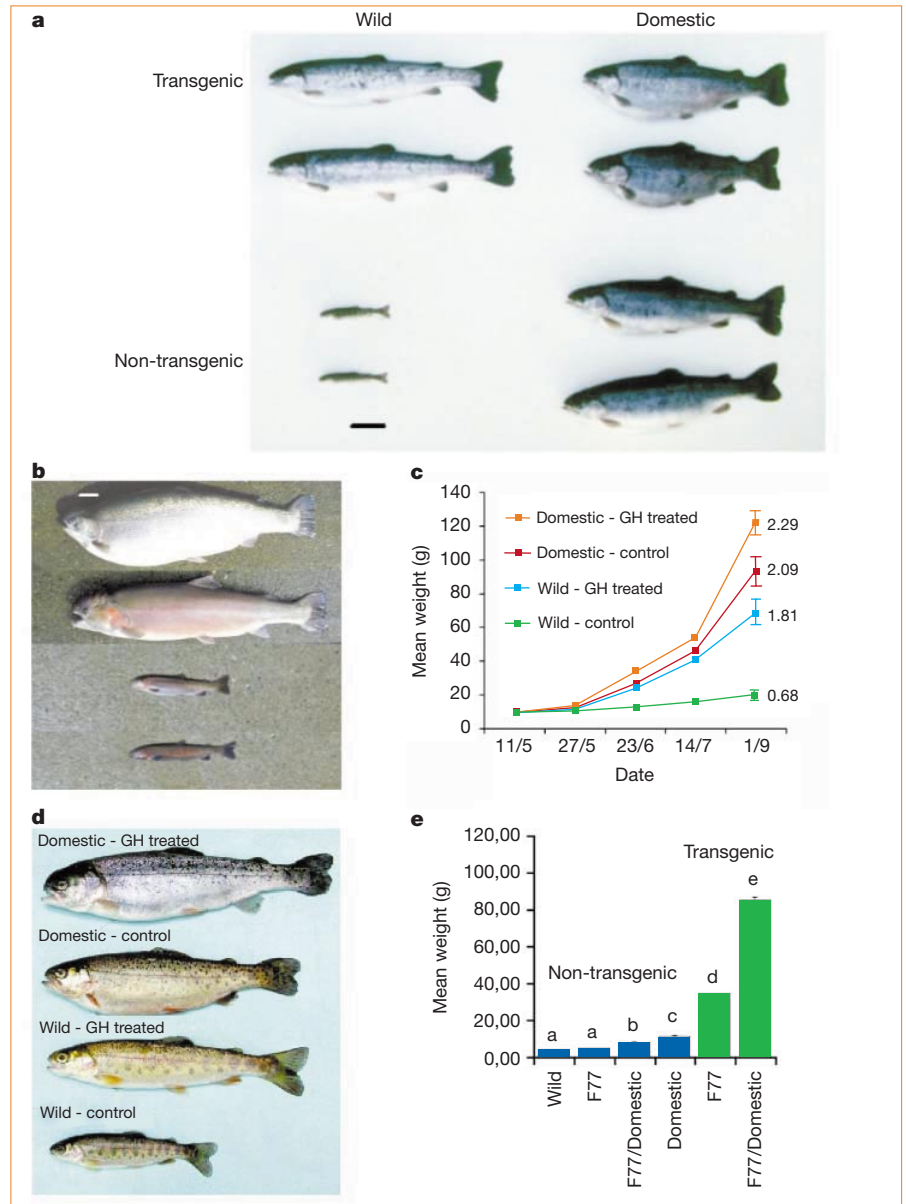


Figure 1 Effect of growth hormone in domestic and wild salmonids. **a**, Pairs of transgenic (top) and non-transgenic (bottom) rainbow trout produced from wild (left) and domesticated (right) strains reared at 8 °C. **b**, Mature wild-strain trout. Top to bottom: cultured transgenic female (14.2 kg), cultured transgenic male (8.2 kg), wild female (171 g), wild male (220 g). Scale bars represent 5 cm. **c**, Growth of domestic and wild rainbow trout strains injected intraperitoneally with 30 μ l of a slow-release formulation (Posilac) of bovine growth hormone⁸. Mean specific growth rates for weight are shown on the right. All groups were significantly different by the end of the trial. **d**, Phenotype of growth-hormone-treated domestic and wild strains of rainbow trout. **e**, Weights (growth in fresh water to 263 days post-fertilization) of non-transgenic (green bars) and transgenic (blue bars) coho salmon in wild, domestic, and wild (F77)/domestic hybrid genetic backgrounds. Strain F77 has a wild genetic background. Statistically distinct groups are indicated by different letters over the bars.

maturation. Cranial abnormalities⁷ detected in transgenic trout were not seen in domestic non-transgenic animals, suggesting that, unlike domestication, transgenesis can affect growth pathways outside the range supported by the homeostatic processes that maintain the fish's normal morphology and viability.

Trout treated with exogenous growth-hormone protein⁸ show similar effects. The size of untreated wild-strain rainbow trout increased with a relatively slow weight-specific growth rate of 0.68% per day, whereas untreated domestic trout grew at more than three times this rate (Fig. 1c). The specific growth rates of hormone-

treated wild-strain trout were enhanced 2.7-fold, whereas domestic trout displayed a much more modest increase of only 9%. Cranial abnormalities and silver body coloration were also seen in only the hormone-treated animals from the domestic strain (Fig. 1d). The fact that domestic trout respond to the exogenous growth-hormone protein but not to *OnMTGH1* transgenesis suggests that 'stronger' gene constructs could be effective, although they might be associated with a higher incidence of abnormalities.

The effect of introducing a growth-hormone gene construct into fish to increase growth rates appears to be dependent on the degree to which earlier enhancement has been achieved by traditional genetic selection. Such effects are likely to be specific for different species, strains and transgenes — in selected mice or in domesticated, rapidly growing farm animals, for example, growth-hormone transgenesis can have little effect on growth or it can induce pathological effects^{9,10}, as we have seen in transgenic salmonids.

Depending on the genetic and physiological basis, not all gains made by selection are likely to be epistatic to the effects of growth-hormone transgenesis, and some might actually prime metabolic pathways to respond to endocrine stimulation. In par-

tially domesticated coho salmon strains, we have observed that domestication and the *OnMTGH1* transgene can work synergistically in hybrid strains to increase overall growth (Fig. 1e). In contrast, growth-hormone constructs that work well in salmon may not work as well in species that grow rapidly⁴. The phenotypic and genetic character of the starting species or strain, as well as the strength of the gene construct, need to be considered when attempting to improve agricultural animal species.

**Robert H. Devlin, Carlo A. Biagi,
Timothy Y. Yesaki, Duane E. Smailus,
John C. Byatt***

Fisheries and Oceans Canada, West Vancouver,

British Columbia V7V 1N6, Canada

**Monsanto Corporation, St Louis, Missouri 63198,*

USA

1. Hershberger, W. K. *et al. Aquaculture* **85**, 1–4 (1990).
2. Du, S. J. *et al. Bio/Technology* **10**, 176–180 (1992).
3. Devlin, R. H. *et al. Nature* **371**, 209–210 (1994).
4. Rahman, M. A., Mak, R., Ayad, H., Smith, A. & Maclean, N. *Transgen. Res.* **7**, 357–369 (1998).
5. Devlin, R. H. in *Transgenic Animals: Generation and Use* (ed. Houdebine, L. M.) 105–117 (Harwood Academic, Amsterdam, 1997).
6. Muir, W. M. & Howard, R. D. *Proc. Natl Acad. Sci. USA* **96**, 13853–13856 (1999).
7. Devlin, R. H., Yesaki, T. Y., Donaldson, E. M. & Hew, C. L. *Aquaculture* **137**, 161–169 (1995).
8. McLean, E. *et al. Aquaculture* **156**, 113–128 (1997).
9. Pursell, V. G. *et al. Science* **244**, 1281–1288 (1989).
10. Eisen, E. J., Murray, J. D. & Schmitt, T. J. *J. Anim. Breed. Genet.* **112**, 401–413 (1995).

has been compressed to a radius $R_c \approx a$. Light is emitted from this region, which is comparable in size to (or smaller than) the wavelength of the light.

According to standard formulae of plasma physics, the photon–matter interaction length is large compared to a (ref. 4). Therefore, the proposed spectrum^{2,5} is not that of a black body, but bremsstrahlung from a thermally ionized plasma. At 17,500 K, the noble gases are only weakly ionized, and the radiation is approximately proportional to the square of the degree of ionization, $e^{-\chi/kT}$, where χ is the ionization potential (25 eV for helium and 12 eV for xenon). Thus, radiation from a uniform He bubble should be less than that from a Xe bubble by more than four orders of magnitude for electron–ion bremsstrahlung, and by two orders of magnitude for electron–neutral bremsstrahlung. But in fact, the observed emission from He is less than that from Xe by only about a factor of 3 (refs 1, 6) — a discrepancy between theory² and experiment⁷ of one to three orders of magnitude⁷.

Another problem with Hilgenfeldt *et al.*'s model is that the theory encapsulated in equations (1) and (2) is applied in a limit where it is not valid⁸. For Rayleigh's equation to apply, the speed of the collapsing bubble wall, $\dot{R}$, must be small compared to the speed of sound in the gas, c_g (that is, the Mach number, $\epsilon \equiv \dot{R}/c_g$, must be much less than 1). But experiments show¹ that $\epsilon_g > 1$, undermining the assumption that the bubble's contents are uniformly compressed².

Although sonoluminescence is arguably nature's most nonlinear oscillator⁹, equation (1) is derived using the linear wave equation to describe the motion of the water. Photographs of high-Mach-number shocks emanating from a bubble¹⁰ show that the physics of sonoluminescence contradicts the simplifications of equations (1) and (2). Löfstedt *et al.*⁸ have also pointed out the inconsistencies involved in applying these equations to a sonoluminescence bubble near R_c .

If the asymptotic expansion leading to equation (1) were applied within its range of validity ($\epsilon_g \ll 1$; $\epsilon \equiv \dot{R}/c \ll 1$), then terms that appear at higher order in ϵ should have a small effect. But this is not the case: adding a typical ϵ^2 term [$\epsilon^2(R^2/\rho)d^2P_g/dR^2$] to the right side of equation (1) now yields a collapse temperature of 9,000 K, and an even greater discrepancy between theory and experiment.

Another problem with Hilgenfeldt *et al.*'s theory² is that it neglects the important role of water vapour^{11,12}. For all noble gases, strongly different sonoluminescence intensities can be observed as the ambient temperature is varied¹. As this effect can occur at fixed bubble size, one concludes that water vapour is a key feature of sonoluminescence^{11,12}.

Cavitation science

Is there a simple theory of sonoluminescence?

An abiding issue in cavitation science is the focusing of energy by the collapse of a gas bubble in water. In particular, one would like to understand the origin of the flash of light ('sonoluminescence') that accompanies bubble collapse¹. Hilgenfeldt *et al.*² have presented a simple explanation for this light emission, based on a hydrodynamic (Rayleigh–Plesset) analysis of bubble dynamics. Here we argue that their model is too simple, on the grounds that it cannot account for some well-established observations and that it involves the application of this hydrodynamic equation outside its range of validity.

Hilgenfeldt *et al.*² calculate the interior temperature of a collapsing gas bubble in water, T_g , from the Rayleigh–Plesset equation for the radius, $R(t)$, of a pulsating bubble:

$$R\ddot{R} + \frac{3}{2}\dot{R}^2 = \frac{1}{\rho}(P_g - P_o - P_a) - \frac{4\eta\dot{R}}{\rho R} - \frac{2\sigma}{\rho R} + \frac{R}{\rho c} \frac{d}{dt}(P_g - P_a) \quad (1)$$

supplemented by adiabatic equations of state for the gas temperature T_g and pres-

sure P_g inside a uniformly compressed bubble:

$$P_g(R) = \frac{P_0 R_0^{3\gamma}}{(R^3 - a^3)^\gamma};$$

$$T_g(R) = \frac{T_0 R_0^{3(\gamma-1)}}{(R^3 - a^3)^{\gamma-1}} \quad (2)$$

where γ is the ratio of heat capacities C_p/C_v ; a is the radius of the gas when compressed to its van der Waals hard core; and ρ , c , σ and η are the density, speed of sound, surface tension and viscosity of water. Equation (2) applies when the motion is sufficiently rapid, when $R \leq R_0$, the equilibrium radius. (For slow pulsations, $T_g \approx T_0$, the ambient temperature.) When an isolated bubble is trapped and driven by a sound field with amplitude $P_a(t) = P_a \cos \omega_a t$, the bubble emits one flash of light with each cycle³.

Although helium and xenon have very different physical properties, the similarity of their sonoluminescence constitutes a litmus test for theories. When dissolved in water at 3 torr partial pressure, these gases can form light-emitting bubbles with $R_0 = 4 \mu\text{m}$ and a maximum radius of $R_m = 30 \mu\text{m}$ at 40 kHz driving frequency. Equations (1) and (2) yield a collapse temperature, T_c , of 17,500 K when the bubble

Attempts to close out research on cavitation luminescence by using equation (1) are not new. In 1966 it was discovered that light emitted by bubbles formed in flow through a Venturi tube¹³ came out in flashes of duration too short to be measured. At the time these were the shortest man-made flashes of light, but this line of research was abandoned when the cavitation 'establishment' declared it uninteresting because all of the results could be parametrized by equation (1) (ref. 14).

If sonoluminescence originates in a transparent plasma¹⁵, and if this plasma is formed from molecules of dissolved gas, such as He or Xe, then the similarity of sonoluminescence from He and Xe suggests the existence of an additional energy-focusing mechanism within the bubble^{3,11} — a mechanism that could create a strongly ionized (nano)plasma.

S. Putterman, P. G. Evans*, G. Vazquez, K. Weninger

Physics Department, University of California, Los Angeles, California 90095, USA
e-mail: puherman@ritva.physics.ucla.edu
*Division of Engineering and Applied Science, Harvard University, Cambridge, Massachusetts 02138, USA

- Barber, B. P. *et al.* *Phys. Rep.* **281**, 65–144 (1997).
- Hilgenfeldt, S., Grossman, S. & Lohse, D. *Nature* **398**, 402–405 (1999).
- Temple, P. R. *Sonoluminescence from the Gas of a Single Bubble*. Thesis, Univ. Vermont (1970).
- Zeldovich, Y. B. & Raizer, Y. P. *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena* Vols 1, 2 (Academic, New York, 1967).
- Wu, C. C. & Roberts, P. H. *Phys. Rev. Lett.* **70**, 3424–3427 (1993).
- Hiller, R. A. *et al.* *Science* **265**, 248–250 (1994).
- Hilgenfeldt, S., Grossman, S. & Lohse, D. *Phys. Fluids* **11**, 1318–1330 (1999).
- Löffstedt, R., Barber, B. P. & Putterman, S. *Phys. Fluids A* **5**, 2911–2928 (1993).
- Puterman, S. *Phys. World* **11**, 38–42 (1998).
- Weninger, K., Evans, P. G. & Putterman, S. *Phys. Rev. E* **61**, 1020–1023 (2000).
- Puterman, S. & Weninger, K. *Annu. Rev. Fluid Mech.* **32**, 445–476 (2000).
- Vazquez, G. & Putterman, S. *Phys. Rev. Lett.* **85**, 3037–3040 (2000).
- Peterson, F. B. & Anderson, T. P. *Phys. Fluids* **10**, 874–879 (1967).
- Hickling, R. *Phys. Fluids* **11**, 1586–1587 (1968).
- Vasquez, G., Camara, C., Putterman, S. & Weninger, K. Los Alamos preprint server 0009057.

Hilgenfeldt et al. reply — The comment by Putterman *et al.* in essence addresses two questions: the role of water vapour inside bubbles that undergo single-bubble sonoluminescence, and the use of Rayleigh–Plesset dynamics to describe collapsing bubbles.

Regarding the first point, water vapour is indeed present in sonoluminescing bubbles^{1–7}: it invades the bubble during its expansion, and at bubble collapse the remaining water vapour and its reaction products (O₂, H₂, ...) contribute to the light emission. We have discussed in detail^{8,9} how the discrepancy between our model and the results for helium bubbles can be explained

by including this effect: as the ionization energy of oxygen and hydrogen is similar to that of argon and higher than that of xenon, the light-emission intensities with additional oxygen and hydrogen atoms in the bubble are hardly different from those for the pure inert gases. In contrast, in the case of helium with its very large ionization energy, the light-emission process is dominated by water and its reaction products. Meanwhile, we have quantitatively included water vapour in our model⁷.

Regarding the second point, it has long been known that the assumptions used to derive the Rayleigh–Plesset equation indeed break down at bubble collapse, and that there are many ways to extend equation (1) of Putterman *et al.* to higher orders in $\dot{R}/c$ (ref. 10). Although the quantitative details depend on which extension is chosen, the trends in the parameter dependences and the orders of magnitude of the energies involved are robust. Equations of types (1) and (2) have provided useful results when applied over the whole oscillation cycle of the bubble, as evidenced in the pioneering work of Gaitan^{11,12} and the later studies of the Putterman group¹³, in which the criticized Rayleigh–Plesset equation was used to fit various parameters to experimental data on $R(t)$.

Finally, we would like to stress that, contrary to what Putterman *et al.* state, equation (2) was never used in our model. Rather, we change γ dynamically, following Prosperetti^{14,15}, which allows for a more realistic heating of the bubble interior.

Our model reproduces the salient features of sonoluminescence light emission with comparatively little computational effort, allowing for direct comparison with experiment over a wide range of control parameters. Far from “closing out” sonoluminescence research, it has promoted quantitative extensions and refinements by others and by ourselves.

S. Hilgenfeldt*†, S. Grossmann‡, D. Lohse†

*Division of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, Massachusetts 02138, USA

†Department of Applied Physics and J. M. Burgers Centre for Fluid Dynamics, University of Twente, PO Box 217, NL-7500 AE Enschede, The Netherlands

e-mail: lohse@tn.utwente.nl

‡Fachbereich Physik der Universität Marburg, Renthof 6, D-35032 Marburg, Germany

- Moss, W., Clarke, D. & Young, D. *Science* **276**, 1398–1401 (1997).
- Moss, W. & Young, D. *J. Acoust. Soc. Am.* **103**, 3076 (1998).
- Moss, W. *et al.* *Phys. Rev. E* **59**, 2986–2992 (1999).
- Storey, B. D. & Szeri, A. *J. Proc. R. Soc. Lond. A* **456**, 1685–1709 (2000).
- Yasui, K. *Phys. Rev. E* **56**, 6750–6760 (1997).
- Yasui, K. *Phys. Rev. E* **59**, 1754–1758 (1999).
- Tögel, R., Gompf, B., Pecha, R. & Lohse, D. *Phys. Rev. Lett.* **85**, 3165–3168 (2000).
- Hilgenfeldt, S. & Lohse, D. *Phys. Rev. Lett.* **82**, 1036–1039 (1999).

- Hilgenfeldt, S., Grossmann, S. & Lohse, D. *Phys. Fluids* **11**, 1318–1330 (1999).
- Prosperetti, A. & Lezzi, A. *J. Fluid Mech.* **168**, 457–478 (1986).
- Gaitan, D. F. *An Experimental Investigation of Acoustic Cavitation in Gaseous Liquids*. Thesis, Univ. Mississippi (1990).
- Gaitan, D. F., Crum, L. A., Roy, R. A. & Church, C. C. *J. Acoust. Soc. Am.* **91**, 3166–3183 (1992).
- Barber, B. P. *et al.* *Phys. Rep.* **281**, 65–143 (1997).
- Prosperetti, A. *J. Acoust. Soc. Am.* **61**, 17–27 (1977).
- Prosperetti, A. *J. Fluid Mech.* **222**, 587–616 (1991).

Genetic imprinting

Urinary odour preferences in mice

Odour cues influence a variety of social activities in mammals, including kin recognition, mate selection, inbreeding avoidance and juvenile dispersal from the natal area^{1–3}. Inbreeding avoidance is particularly evident across the mammalian phyla because inbreeding can cause a reduction in fitness⁴. Here we show that the attraction of mice to the urinary odours of other mice is subject to a ‘parent-of-origin’ effect⁵ which causes both males and females to prefer the odour of urine from mice of an unrelated strain to that of urine from mice of the same strain as their mothers.

As the genes of inbred strains of mice are homozygous at nearly all loci, reciprocal crosses between two independent strains will produce first-generation (F₁) offspring with the same complement of genes. However, each cross will differ in their expression of those genes that are subject to a parent-of-origin effect and, if these are polymorphic, the offspring may exhibit different phenotypes. To exclude the possibility that odour preference might be influenced by familial imprinting⁶, we used mice derived by embryo transfer to genetically unrelated foster mothers.

We tested F₁ mice to see whether they had any preference for urine from either maternal- or paternal-strain females compared with urine from non-related female controls (BALB/c mice). The animals were given a choice of two urine samples with different odours, placed in the end chambers of a three-chambered arena. We measured the time each mouse spent in each end chamber over a two-minute test period. Male and female (CBA/Ca × C57Bl/6)F₁ mice (maternal strain is written first) showed a significant preference for the control urine ($P < 0.05$) over urine from the maternal strain (CBA/Ca) (Fig. 1a). This cross did not show any preference between the control and the paternal-strain (C57Bl/6) urine.

The reciprocal cross (C57Bl/6 × CBA/Ca) showed the same pattern of preference with respect to parental origin, but the opposite pattern of preference with respect to genotype (Fig. 1b). Males and females of this cross preferred the control odour to that of

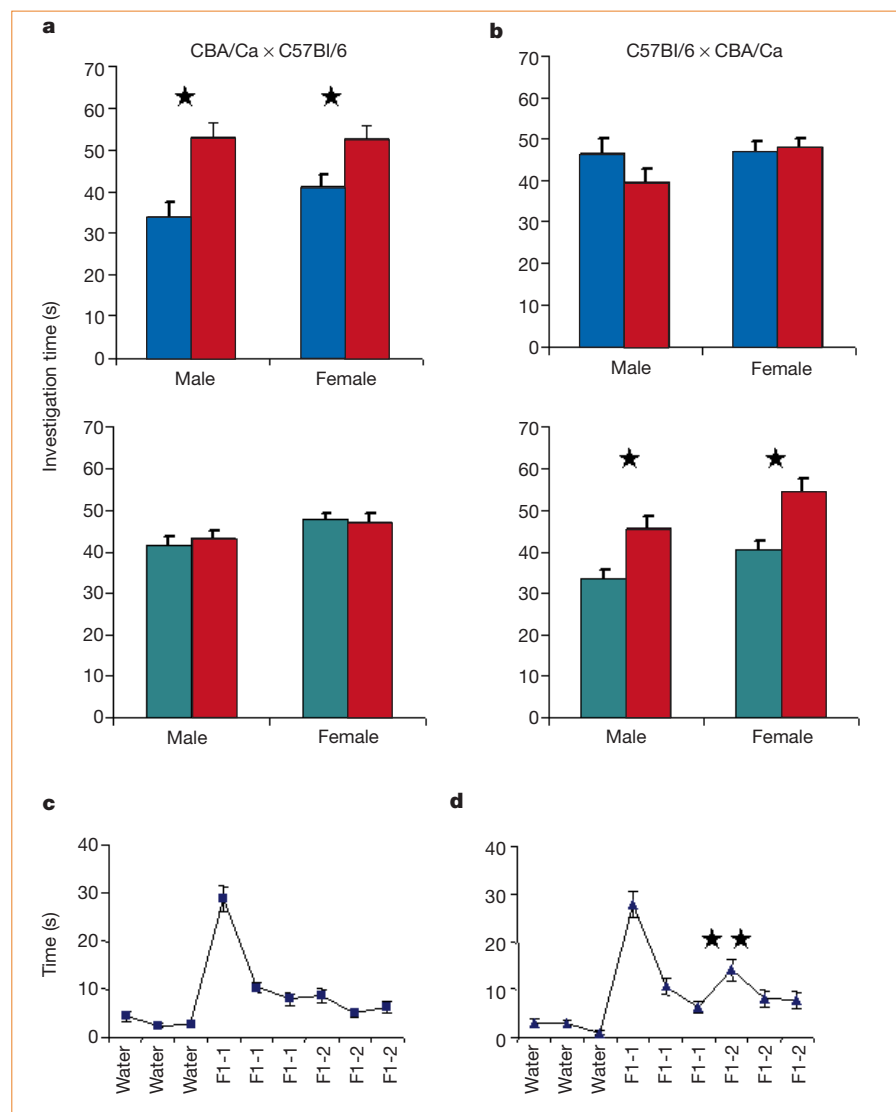


Figure 1 Odour preference of reciprocal-cross F_1 males and females for female urinary odours. Test urine from different mouse strains: blue, CBA/Ca; red, BALB/c (unrelated controls); green, C57Bl/6. **a**, (CBA/Ca × C57Bl/6) F_1 males and females show a preference for BALB/c female urine compared to CBA/Ca female urine, but no preference for BALB/c over C57Bl/6 female urine. **b**, (C57Bl/6 × CBA/Ca) F_1 males and females show no preference between CBA/Ca and BALB/c female urinary odours, but favour BALB/c over C57Bl/6 female urine. Mean investigation times $\pm$ s.e.m. are shown; asterisks indicate $P < 0.05$ (Wilcoxon test). **c**, In habituation–dishabituation tests, BALB/c females cannot distinguish between reciprocal-cross F_1 male urines (no significant increase in investigation time for F1-1 and F1-2). **d**, The same mice can distinguish between C57Bl/6 and CBA/Ca urine. Mean investigation times $\pm$ s.e.m. are shown; asterisks indicate $P < 0.01$ (t -test). Further details of experimental procedures are available from the authors.

the maternal-strain (C57Bl/6) urine ($P < 0.05$), but showed no preference between the control and paternal-strain CBA/Ca odours.

Reciprocal F_1 females were tested in the same way for which male urinary odours they favoured, but none showed any significant preference in either of the two tests. To ensure that this lack of a preference was not due to any impairment in olfactory function, we also tested these mice for their ability to distinguish between urine from C57Bl/6 and CBA/Ca males using a habituation–dishabituation test⁷. The females were found to have a normally functioning olfactory system, and were clearly able to distinguish between the two odour types ($P < 0.01$).

As our F_1 mice were derived by embryo transfer to unrelated foster mothers, the parent-of-origin effect on their odour preferences could not have been due to prior exposure to odour cues from their genetic parents. The odour preferences shown by reciprocal F_1 mice are probably ‘hard-wired’ by genes subject to parent-of-origin effects. It is unclear how this odour preference is determined in the reciprocal mice and at what level the effects are exerted. For instance, the genetic parent might influence the odour produced by the F_1 mice themselves. If genes subject to parent-of-origin effects influence normal odour type, then reciprocal F_1 mice should have different representations of their own odours, which in turn could affect their odour preference⁸.

To test this idea, we examined the ability of BALB/c female mice to distinguish between the odours of reciprocal-cross F_1 males using the habituation–dishabituation paradigm. These females could not distinguish between the urines from reciprocal F_1 males (Fig. 1c; $P = 0.35$), but could readily distinguish urine from the individual parental strains (C57Bl/6 and CBA/Ca) (Fig. 1d; $P < 0.01$).

Our findings indicate that the parent-of-origin effect on odour preference in the offspring is not an acquired trait and that it may be due to differences in the functioning of the olfactory system in the reciprocal F_1 types, either at the level of perception or information processing. As these effects occur in mice of both sexes, they are probably due to autosomal imprinting and are not sex-linked.

Genes subject to a parent-of-origin effect that directly influences olfactory-related behaviour may have evolved to promote mechanisms of inbreeding avoidance, possibly by diminishing preference for maternal urine odour and thereby encouraging dispersal. If such a genetic system is to operate in outbred populations, the genes that determine odour and those subject to parent-of-origin effects acting on the perception of odours may need to be in linkage disequilibrium, for which there is precedent in both mice and humans^{9,10}.

Anthony R. Isles[†], Michael J. Baum[‡],
Dan Ma[†], Eric B. Keverne[†],
Nicholas D. Allen^{*}

^{*}Laboratory of Cognitive and Developmental Neuroscience, Babraham Institute, Babraham CB2 4AT, UK

e-mail: nick.allen@bbsrc.ac.uk

[†]Sub-Department of Animal Behaviour, University of Cambridge, Madingley CB3 8AA, UK

[‡]Department of Biology, Boston University, 5 Cummington Street, Boston, Massachusetts 02215, USA

- Waldman, B., Frumhoff, P. C. & Sherman, P. W. *Trends Ecol. Evol.* **3**, 8–13 (1988).
- Moore, J. & Ali, R. *Anim. Behav.* **32**, 94–112 (1984).
- Barnard, C. J. & Fitzsimons, J. *Anim. Behav.* **38**, 35–40 (1989).
- Meagher, S., Penn, D. J. & Potts, W. K. *Proc. Natl Acad. Sci. USA* **97**, 3324–3329 (2000).
- Barlow, D. P. *Science* **270**, 1610–1613 (1995).
- Yamazaki, K. *et al. Science* **240**, 1331–1332 (1988).
- Sundberg, H., Doving, K., Novikov, S. & Ursin, H. *Behav. Neural Biol.* **34**, 113–119 (1982).
- Mateo, J. M. & Johnston, R. E. *Proc. R. Soc. Lond. B* **267**, 695–700 (2000).
- Amadou, C. *et al. Immunol. Rev.* **167**, 211–221 (1999).
- Fan, W., Liu, Y. C., Parimoo, S. & Weissman, S. M. *Genomics* **27**, 119–123 (1995).

erratum

Watching fights raises fish hormone levels

R. F. Oliveira, M. Lopes, L. A. Carneiro, A. V. M. Canário
Nature **409**, 475 (2001)

The Mann–Whitney test P values given in Fig. 1 were incorrect. In Fig. 1a, these should have been (from left to right) $P < 0.05$, $P < 0.01$ and $P < 0.01$; and $P < 0.05$ in Fig. 1b.

Nuclear fission modes and fragment mass asymmetries in a five-dimensional deformation space

P. Möller*, D. G. Madland*, A. J. Sierk* & A. Iwamoto†

* Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

† Department of Materials Science, Japan Atomic Energy Research Institute, Tokai-mura, Naka-gun, Ibaraki, 319-1195 Japan

Nuclei undergoing fission can be described by a multi-dimensional potential-energy surface that guides the nuclear shape evolution—from the ground state, through intermediate saddle points and finally to the configurations of separated fission fragments. Until now, calculations have lacked adequate exploration of the shape parameterization of sufficient dimensionality to yield features in the potential-energy surface (such as multiple minima, valleys, saddle points and ridges) that correspond to characteristic observables of the fission process. Here we calculate and analyse five-dimensional potential-energy landscapes based on a grid of 2,610,885 deformation points. We find that observed fission features—such as the distributions of fission fragment mass and kinetic energy, and the different energy thresholds for symmetric and asymmetric fission—are very closely related to topological features in the calculated five-dimensional energy landscapes.

When a heavy nucleus divides into two fragments in nuclear fission, two key questions about the process have challenged researchers since the discovery of fission more than 60 years ago. First, what is the threshold energy for the reaction and, second, what shape changes are involved in the transition from a single nuclear system to two separated daughter fragment nuclei? These two questions are intimately connected. The potential energy of a nucleus as a function of shape defines a landscape in a multi-dimensional deformation space. In this landscape, the energy of the lowest mountain passes, or saddle points, connecting the nuclear ground state with the region corresponding to separated fragments, represents the threshold energy of the fission process.

Previous theories and results

The first theory of fission, put forward in 1939 by Meitner and Frisch¹ and Bohr and Wheeler², explained the break-up of uranium into two lighter fragments of roughly equal size with a model involving a charged liquid drop with surface tension. This break-up had been observed just months earlier by Hahn and Strassmann³ in the reaction $n + \text{U}$. In such a macroscopic model the drop becomes increasingly less stable with respect to deformation when the atomic number Z increases, and at $Z \approx 100$ stability is completely lost. For slightly lighter actinide nuclei the fission barrier between the ground-state shape and the separated-fragment configuration is sufficiently low that spontaneous fission, due to quantum-mechanical penetration of the fission barrier, occurs with measurable probability. Fission may also be induced by exciting the nucleus to energies above the barrier energy. For example, a thermal neutron incident upon ^{235}U imparts sufficient energy to excite the compound nucleus ^{236}U above the barrier.

In a pioneering use of the first electronic digital computer ENIAC in 1947, Frankel and Metropolis⁴ explored some key aspects of the macroscopic liquid-drop model. In particular, they determined the shapes of fissioning nuclei at the saddle-point thresholds. In the 1960s a greatly improved model for the nuclear potential energy as a function of shape emerged. In this macroscopic–microscopic model, the potential energy is the sum of shape-dependent macroscopic (liquid-drop) and microscopic (single-particle) terms. Over the past 30 years the model has provided considerable insight into

nuclear structure. For example, improved descriptions of fission-isomeric states and mass-asymmetric fission saddle points have been obtained and nuclear masses are calculated for nuclei throughout the periodic system to an average root-mean-square (r.m.s.) accuracy of about 0.7 MeV (refs 5–13).

The microscopic energy, calculated by following the method developed by Strutinsky^{5,6} and representing a modification to the slowly varying macroscopic ‘liquid drop’ energy of the nucleus, is due to the presence of quantum-mechanical structure in the nucleus. For chemical properties it is well known that certain elements, the noble gases, are unusually stable and non-reactive, due to particularly stable electron configurations, which occur for 2, 10, 18, 36, 54 and 86 electrons, corresponding to He, Ne, Ar, Kr, Xe and Rn. At these numbers large gaps occur in the energy-level spectrum corresponding to the individual electron orbitals. In nuclei, quantum-mechanical laws give rise to increased stability at similar gaps corresponding to nuclear ‘magic numbers’ for protons and neutrons. Because of differences between the purely electromagnetic forces determining the electron levels and the nuclear forces, the first few ‘magic numbers’ for spherical nuclei are 2, 8, 20, 28, 50 and 82 for both protons and neutrons. However, although the effects are largest for the ‘magic’ nuclei, microscopic corrections to the simple liquid-drop model of nuclei occur to some degree in all nuclei. Strutinsky^{5,6} generalized the concept of magic numbers so that a precise, well specified microscopic (shell) correction to the liquid-drop energy can be calculated for any shape and any particle number. It is the shell-correction part of the total nuclear energy that is responsible for the multiple valleys, minima, saddle points and peaks that appear in general multi-dimensional potential-energy surfaces as functions of nuclear shape. In such surfaces it has been demonstrated, for some nuclei in limited shape parameterizations, that ‘shells’ or regions of unusually low energy corresponding to magic numbers in separated fission fragments manifest themselves as deep valleys in the nuclear potential energy surface long before division into separate daughter fragments^{14–17}. However, to what extent ridges stabilize these valleys was unclear in these previous, lower dimensional, schematic calculations.

In the past, fission properties have often been correlated with models of the binding energy of separated fission fragments, and of valleys inside the point of contact. However, the valleys by

themselves do not determine the final state of a fissioning nucleus. Final states corresponding to three or more fragments are in many cases energetically more favoured than are states of two final fission fragments. In these cases the nucleus nonetheless divides into only two fragments. This occurs because the barrier between the ground state and the binary fission valley favours such divisions and a ridge separates the binary from the ternary valley, although dynamical effects may also affect the division. Here we examine what saddles connect the ground state to the various two-fragment valleys that emerge in the later stages of the fission process, and what are the heights of the ridges that allow or prevent movement between the valleys.

Unsolved mysteries

With our new approach we are now able to explain a number of previously unresolved, very important characteristic features of fission. These include:

(1) Nuclei just below the actinide region close to ^{228}Ra exhibit two fission modes^{18–21}. Figure 1 illustrates experimental data obtained for ^{227}Ra in ref. 19. In one mode, with the lower threshold energy, the fragment mass distribution is asymmetric and the total fragment kinetic energy is about 10 MeV higher than in the other, symmetric, mode¹⁹. At certain excitation energies these two modes lead to a striking three-peaked structure of the fragment-mass yield curve. From the totality of the experimental data the authors of ref. 19 conclude: “Thus it seems that after the gross determination of the symmetric or asymmetric character of fission made already at the barrier, the two components follow a different path with no or little overlap in the development from the barrier to the scission configuration.”

(2) Nuclei at the upper end of the actinide region exhibit sudden changes with nucleon number in fission properties and sometimes display a two-mode character in the same nucleus. For example, the fragment mass distribution changes abruptly from mass asymmetric for ^{256}Fm to mass symmetric for ^{258}Fm along with a correlated increase in the fragment total kinetic energy (TKE) by about 35 MeV. But ^{258}Fm also exhibits the asymmetric mode with lower TKE with a small probability: fission of such nuclei is characterized as bimodal.

(3) Throughout the actinide region below Fm, nuclei near the line of β -stability in spontaneous or low-energy induced fission divide into a heavy fragment with a mass close to 140 and a light fragment with a mass that shifts with the total mass of the fissioning nucleus. An

example of a typical fission-fragment charge (mass) distribution is shown in Fig. 2. In our new strategy for applying the macroscopic–microscopic method to higher-dimensional spaces, all of these observed fission phenomena can be understood in terms of nuclear potential-energy surfaces calculated with five appropriately chosen nuclear shape degrees of freedom.

New approach

Since the early 1970s (refs 5–11), there has been no major improvement in the description of the fission potential-energy landscape, although many calculations based on 1,000 or so points in deformation space have been presented. We have learned that to describe properly the evolution of a single nuclear shape into two fragments of different mass and deformation—for example, one spherical ^{132}Sn -like fragment and one deformed fragment with mass number A near 100—at least five independent shape parameters are required. Earlier, approaches such as self-consistent Hartree–Fock calculations (see discussion in ref. 22) constrained with respect to one variable were sometimes thought to take into account automatically all additional shape degrees of freedom. In other approaches, also called multi-dimensional, the calculated energy was displayed as energy contour maps in terms of two shape degrees of freedom where the energy at each point was minimized with respect to additional shape coordinates. Neither approach results in a correct description of the structure of the full, multi-dimensional fission potential-energy surface. In fact, they are at least as inexact as two-dimensional calculations, as is discussed in more detail in ref. 22. To establish the structure of multi-dimensional spaces it is necessary to calculate the energies corresponding to all physically possible combinations of the deformation parameters, which leads to multi-million-point potential-energy spaces rather than only 1,000 or so points.

We investigate in detail such a space incorporating 2,610,885 points defining a five-dimensional shape-coordinate grid. For the potential energy we use the macroscopic–microscopic finite-range liquid-drop model with shape-dependent Wigner and A^0 terms as defined in refs 12 and 16. Specifically the five shape coordinates are: (1) elongation, expressed in terms of the charge quadrupole moment Q_2 ; (2) neck diameter d ; (3) left nascent-fragment deformation ϵ_{f1} ; (4) right nascent-fragment deformation ϵ_{f2} ; and (5) mass asymmetry α_g , as illustrated in Fig. 3. The charge quadrupole moment Q_2 is given as that of ^{240}Pu with the same shape as the nucleus being considered, so that the nuclear size effect is

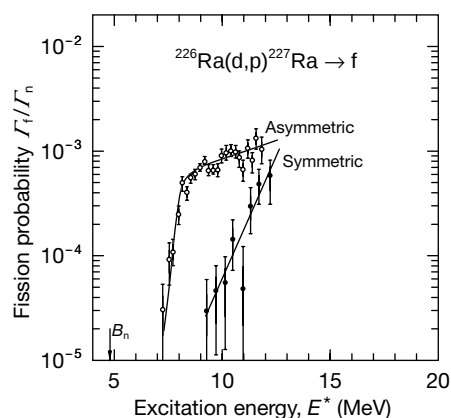


Figure 1 Asymmetric and symmetric fission probabilities for ^{227}Ra as functions of the excitation energy in the fissioning nucleus. Here d denotes a deuteron, p a proton, B_n the neutron binding energy, and Γ_n and Γ_f are proportional to the neutron-emission and fission probabilities, respectively. The data show that two distinct fission modes coexist in this nucleus, namely one asymmetric mode and one symmetric mode, the latter with a 1–2 MeV higher threshold energy. The figure is based on a figure in ref. 19.

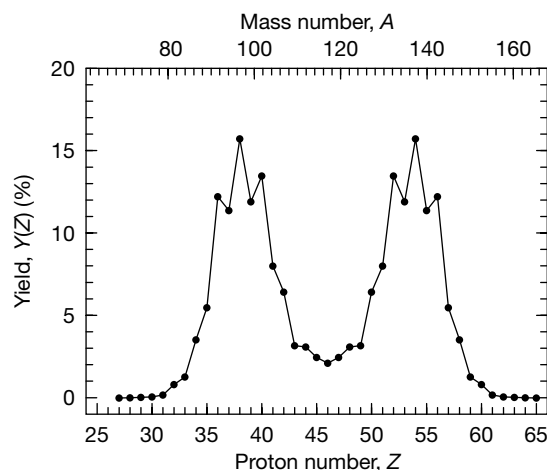


Figure 2 Nuclear charge yield in electromagnetic-induced fission of ^{234}U from ref. 28. The data are converted to a mass–yield distribution before neutron emission (top axis) by assuming that the proton/neutron ratio Z/N is the same in each of the two fission fragments as in the original nucleus.

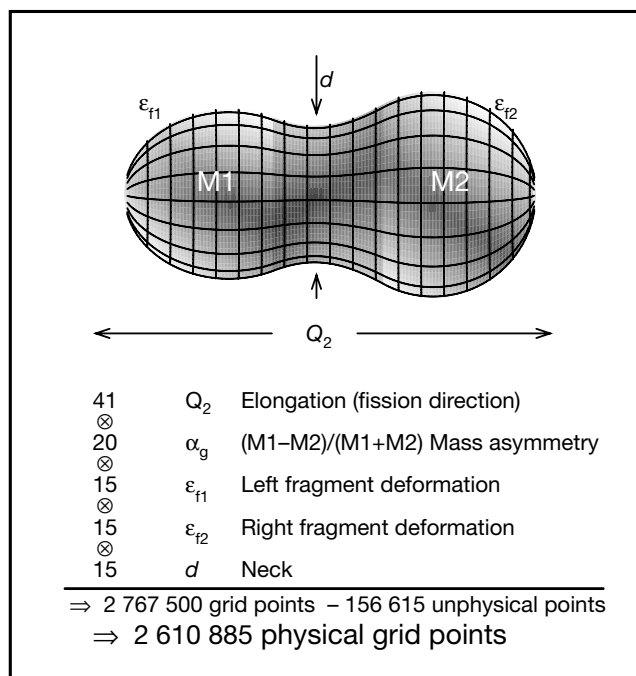


Figure 3 Five-dimensional shape parameterization used in our potential-energy calculation. Different shades of grey indicate the three different quadratic surfaces of the shape parameterization used in our calculation. The first derivative is continuous at the intersections of the surfaces. The crossed circles indicate that the five one-dimensional spaces of each coordinate combine (multiply together) to yield a five-dimensional space with 2,767,500 grid points. However, shapes corresponding to certain quadrupole moments do not exist for specific combinations of the other shape parameters. For example, zero quadrupole moment cannot be realized for shapes with very deformed ends. In our grid there exist 156,615 such ‘unphysical’ points. Thus, we are left with 2,610,885 shapes for which we actually calculate the potential energy. We have closely spaced the asymmetry coordinate so that we will be able to identify favourable saddle-point shapes, close to fragment magic proton and neutron numbers, that may not appear in a more sparsely spaced grid. For ^{240}Pu the spacing corresponds to a change of 2.4 units in the nascent fragment mass numbers.

eliminated. The end-body masses (or, equivalently, volumes) M_1 and M_2 are the masses of the left and right nascent fragments were they completed to closed shapes. The nascent fragments are partial spheroids whose deformations we characterize by Nilsson’s quadrupole ϵ parameter⁷. They are smoothly joined by a partial spheroid or hyperboloid of revolution. Our approach is based on an established realistic microscopic interaction^{12,23}, and the energy behaves properly as the shape evolves from the one limit of a single shape corresponding to the nuclear ground state to the other limit of scission configurations corresponding to two touching daughter fission fragments. By ‘properly’ we mean that the model is formulated so that for a touching-fragment configuration we obtain the same energy whether the energy is calculated as that of a single, very deformed nucleus or as that of two touching nuclei¹⁶.

We identify significant structures in the calculated five-dimensional potential-energy space by considering imaginary water flows²⁴ in five dimensions²². The threshold saddle-point energy and the corresponding shape are found by tracing this imaginary water flow across saddle points when various minima are gradually filled with water.

Given the determination of the threshold energies for fission and the nuclear shapes corresponding to the saddle-point locations, we turn to the second key question: what are the shape changes involved in the transition from a single nucleus to two separated daughter fragment nuclei? Do structures exist in the potential-energy surface that lead to multi-mode fission such as that of the well known three-peaked mass distribution in ^{228}Ra fission¹⁹? To look for such structures we ask whether there are valleys of distinctly different character running in the fission direction of increasing Q_2 . That is, for ten or more fixed Q_2 values beyond the outer saddle region, we determine all minima in the remaining four-dimensional space of the two fragment deformations, neck size and mass asymmetry. We find that there are usually two (but sometimes more) distinct valleys in the region beyond the second saddle region, one corresponding to a mass asymmetry α_g of about $[140 - (A - 140)]/A$ and one corresponding to mass symmetry $\alpha_g = 0$. To understand the significance of these valleys it is necessary to determine more details about their interconnections in the five-dimensional deformation-energy space.

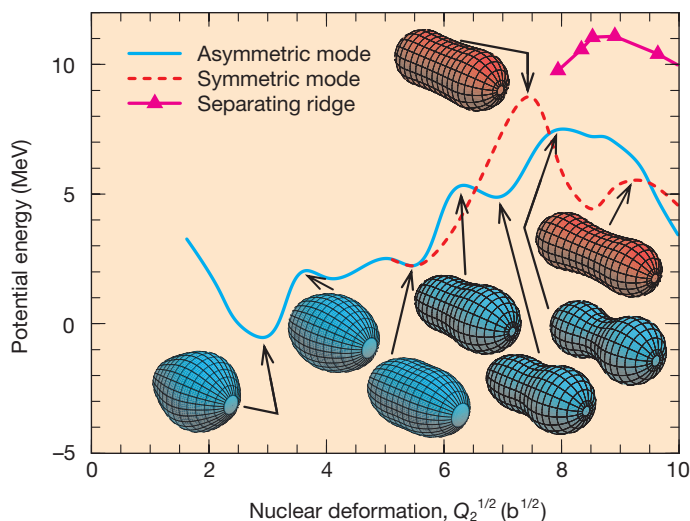


Figure 4 Calculated potential-energy valleys and ridges and corresponding nuclear shapes for ^{228}Ra . Two fission paths exist: one asymmetric path and one symmetric path. The symmetric path has a higher fission saddle point and the more elongated shapes in the valley beyond the saddle point indicate that total fragment kinetic energies in the

symmetric mode are lower than in the asymmetric mode. For excitation energies just above the symmetric saddle the ridge separating the two valleys is high enough to keep the two modes well separated.

Variations of the flooding algorithm allow us to determine that separate saddle points provide entries to the two valleys and the respective energies of these saddle points. Once the lowest saddle has been determined we may block the water flow across this saddle by building an imaginary dam across the saddle region. We can also totally block the water flow beyond a selected maximum Q_2 . This prevents water from flowing down one valley and up 'the back way' into the other valley. To determine the height of the ridge between the two valleys along their entire length, for each fixed Q_2 , we study the remaining four-dimensional space in which the two valleys correspond to two minima and the ridge to the saddle separating them. We use the flooding algorithm in four dimensions to localize this saddle/ridge.

Mysteries resolved

As examples of structures we have found in the calculated five-dimensional potential-energy surfaces we show in Figs 4 and 5 some fission-valley and separating-ridge features obtained for ^{228}Ra and ^{234}U . Asymmetric fission dominates in both cases, because the barrier leading into the asymmetric valley is the lower one, in agreement with experiment. At the saddle points leading to the asymmetric fission valleys the shell correction is already about half of the one calculated for the spherical ground-state shape of ^{132}Sn , whereas the shell correction at the saddle points leading to the symmetric fission valleys is essentially zero. The calculated bi-valley structure of the potential-energy surface leads to the observed bimodal fission features in this region of nuclei^{19,21}. The high ridge separating the two valleys for ^{228}Ra peaks at 2.47 MeV above the

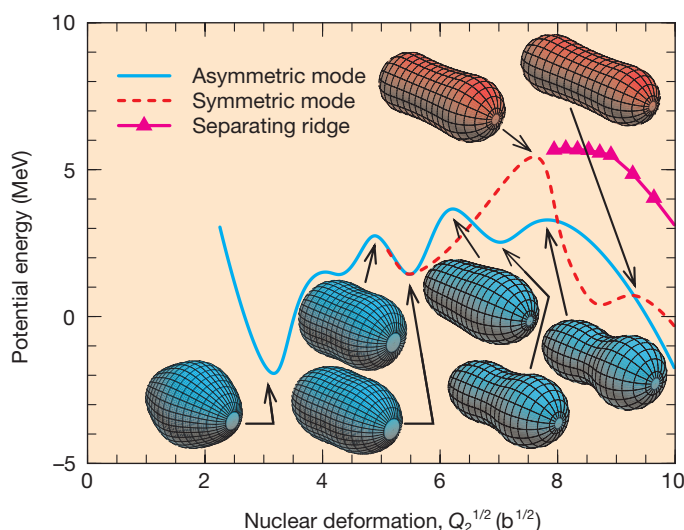


Figure 5 Calculated potential-energy valleys and ridges and corresponding nuclear shapes for ^{234}U . Two fission paths exist: one asymmetric path and one symmetric path. The symmetric path has a higher fission saddle point and the more elongated shapes in

the valley beyond the saddle point indicate that total fragment kinetic energies in the symmetric mode are lower than in the asymmetric mode. The ridge separating the two valleys is certainly not high enough to permit two well-separated modes to evolve.

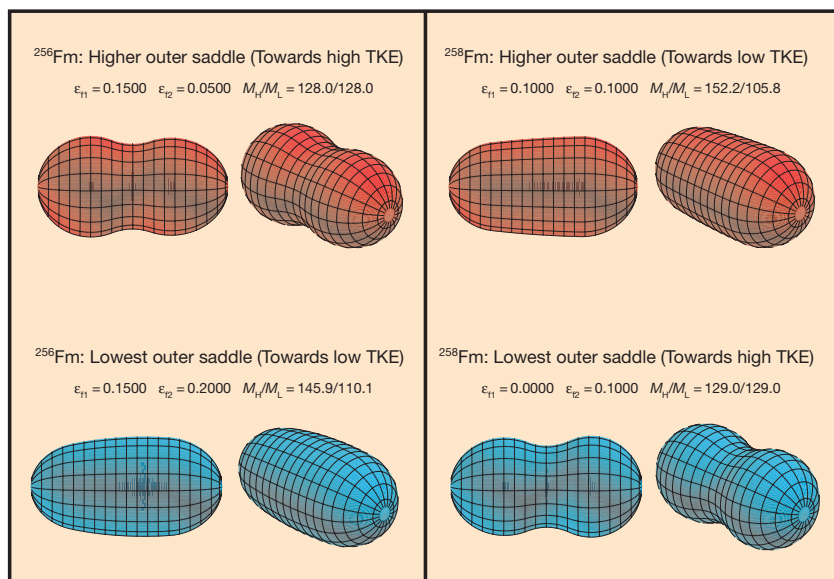


Figure 6 Several saddle-point shapes for ^{256}Fm and ^{258}Fm calculated on the grid in ref. 22. Two views are plotted for each calculated shape; a side view and a view from an angle. For ^{258}Fm the lowest saddle-point energy corresponds to a compact, mass-symmetric shape configuration, which with its nearly spherical nascent fragments corresponds to the experimental observation of high fission-fragment kinetic energies and

mass-symmetric fission. An observed weak mode of low-kinetic-energy fission corresponds to fission in a valley accessible across the higher calculated saddle in the top right part of the figure. For ^{256}Fm the calculated heights of these two saddles are reversed, favouring both low fission-fragment kinetic energies and fragment mass asymmetry, in agreement with the experimental observations. TKE stands for total kinetic energy.

entrance saddle to the symmetric valley. At low excitation energies it therefore keeps the mass-symmetric and mass-asymmetric modes well separated until scission, whereas for ^{234}U the lower separating ridge, at almost the same energy as the entrance saddle to the symmetric valley, allows the symmetric component to partially or completely revert back to the asymmetric valley before scission. The elongated shapes obtained in the symmetric valley are consistent with the lower fragment kinetic energies observed in the symmetric fission mode relative to the asymmetric mode for which we obtain more compact shapes. The fragment kinetic energies arise from Coulomb repulsion which mainly becomes effective just after division into separated fragments. The centres of charge of more elongated touching fragments are farther apart than for compact touching fragments. High fragment kinetic energies therefore indicate compact scission configurations, whereas low fragment kinetic energies indicate more elongated scission configurations.

Nuclei in the region near ^{258}Fm also exhibit bimodal fission features²⁵. We have earlier tentatively identified bimodal structures in calculated two-dimensional potential-energy surfaces^{15,16}, but only now have we verified that these interpretations remain valid when the calculation is extended from two to five dimensions. For ^{256}Fm and ^{258}Fm we find the two distinct classes of saddle points shown in Fig. 6. For ^{256}Fm the shape of the lowest saddle indicates it

corresponds to normal, low-TKE fission similar to what is observed in fission of lighter actinides. However, another saddle point exists, which we calculate to be 0.30 MeV higher than the lower saddle point. This saddle-point shape shows that it corresponds to a path leading to symmetric fission with compact scission configurations and higher fragment kinetic energies. For ^{258}Fm the latter type of saddle point is the lowest saddle point. Thus, we reproduce the experimentally observed transition point between asymmetric low-TKE fission and symmetric high-TKE fission²⁵.

For elongations between the outer fission-barrier saddle point and scission we can unambiguously identify a mass-asymmetric valley for most actinide nuclei. We have defined the deformation-grid mass-asymmetry parameter α_g (compare Fig. 3) as

$$\alpha_g = \frac{M_1 - M_2}{M_1 + M_2} \quad (1)$$

where M_1 and M_2 are the volumes inside the end-body quadratic surfaces, were they completed to form closed-surface spheroids. Sufficiently far along in the mass-asymmetric valley, say at $Q_2 = 64 \text{ b}$, where 1 b is 1 barn, corresponding to 10^{-28} m^2 , the fissioning nuclei always have well-developed necks (see Figs 4 and 5). Therefore it is meaningful to compare the mass asymmetry of the non-separated shape in the mass-asymmetric valley to the final heavy and light fragment masses M_H and M_L .

Although M_1 and M_2 for shapes with well-developed necks can be expected to be close to the final fragment masses we cannot directly compare M_1 and M_2 to the observed fission fragment masses M_H and M_L because the former do not quite sum up to the total nuclear volume or mass A . However, by scaling M_1 and M_2 so that their sum adds up to the total mass number A , we can directly compare the mass asymmetry of the valley shape to the observed heavy and light fragment masses. We obtain trivially

$$M_L^{\text{calc}} = r_s M_1 = A \frac{1 + \alpha_g}{2}, \quad M_H^{\text{calc}} = r_s M_2 = A \frac{1 - \alpha_g}{2} \quad (2)$$

$$\text{and } r_s = \frac{A}{M_1 + M_2}$$

where r_s is a scaling factor for a nucleus with A nucleons. The scaling is equivalent to a redistribution of the mass in the neck region to the left and right masses in proportion to their respective volumes. The amount of matter in this 'gedanken' redistribution is quite small, about 10–20 nucleons.

We use the above definitions in Fig. 7 to compare calculated heavy and light fission-fragment masses, based on valley properties at $Q_2 = 99 \text{ b}$, to experimental data for the mean positions of the heavy- and light-mass peaks in fragment-mass distributions^{26–28} for a range of even isotopes of Th, U, Pu, Cm, Cf and Fm. Our calculated results are in excellent agreement with experimental data, with a mean deviation of only 3.0 nucleons. Specifically, we find that the calculated heavy-fragment mass is roughly constant and close to $A = 140$ for all elements and all isotopes considered. Consequently the mass of the light fragment, which contains the remainder of the mass of the fissioning nucleus, depends on the mass number of the fissioning system, again in excellent agreement with experimental data.

Discussion

The new theoretical results on nuclear fission presented here have been obtained in our standard nuclear-structure model^{29,30} that has also been applied to the calculation of nuclear masses¹², heavy-ion interaction barriers³¹, nuclear β -decay³², and nucleosynthesis in stellar environments³³, with no change in the model or its parameters relative to their 1992 definitions in ref. 12. We have used the FRLDM (1992) version¹² of the potential-energy model, rather than the FRDM (1992) version¹² because the latter is unsuitable for shapes with well-developed necks. We have calculated five-dimensional potential-energy landscapes for 138 even-even nuclei from

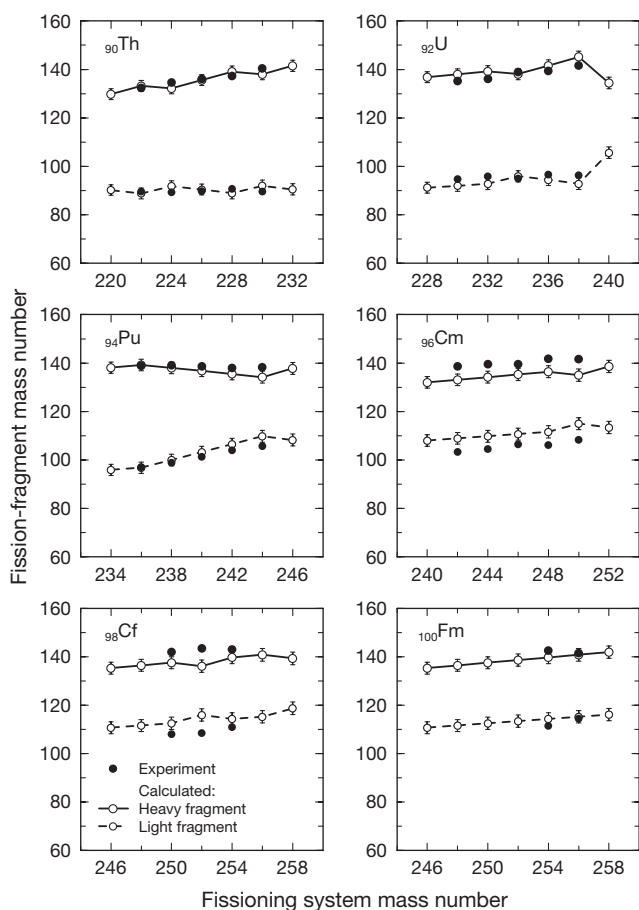


Figure 7 Calculated (white circles) and measured (black circles)^{26–28} average mass division in asymmetric fission for sequences of even isotopes of Th, U, Pu, Cm, Cf, and Fm. The error bars on the calculated points correspond to the spacing of mass asymmetry values on the multidimensional shape-coordinate grid. The data are for spontaneous fission when they are available, otherwise data for low-energy induced fission are used. The results reproduce the experimental observation of a heavy fragment at mass number $A \approx 140$ and a light fragment with mass corresponding to the remainder of the original nucleus. However, deviations from this rule of thumb are also reproduced by the calculations.

Pb to Fm. Our analysis of these landscapes, of which only some examples have been discussed in detail above, allows us to draw the following conclusions:

- (1) Multiple fission paths are found for most nuclei.
- (2) For radium and light actinide nuclei two paths dominate: one mass asymmetric and one mass symmetric. These paths correspond to different fission modes, such as those illustrated in Fig. 1.
- (3) The difference in energy between the symmetric and asymmetric saddle points for Ra and light actinide nuclei in our calculated potential-energy surfaces is 1–2 MeV, which is consistent with the experimentally inferred differences in threshold energies for these two modes^{19–21,34}.
- (4) For ²²⁸Ra and nearby nuclei we find that the symmetric and asymmetric fission paths are well separated by a high ridge from saddle to scission. Thus, our five-dimensional calculations have verified the experimental conclusion reached in ref. 19 that at low excitation energies two fission paths exist with little or no overlap.
- (5) From experimental observations of fission of elements lighter than Fm it is deduced that the average kinetic energy is 10–15 MeV higher for the asymmetric mode than for the symmetric mode^{18,19,35}. The shape differences we calculate for nuclei evolving in the mass-asymmetric and mass-symmetric valleys are qualitatively consistent with these differences in the kinetic energies of the two modes.
- (6) The saddle-point shapes and energies for ²⁵⁶Fm and ²⁵⁸Fm are consistent with the experimentally observed transition between asymmetric low-TKE fission and symmetric high-TKE fission that occurs between ²⁵⁶Fm and ²⁵⁸Fm.
- (7) The long-observed mass split in mass-asymmetric fission with a roughly constant heavy fragment mass near $A = 140$ is convincingly reproduced in our calculations.

As early as 1950 Meitner³⁶ suggested an interpretation of the observed mass divisions in fission in terms of fragment shell structure. We have shown here that it is not only the mass division that is influenced by shell structure. Shell structure also creates different modes of fission, each of which has its characteristic saddle-point energy, mass division, and kinetic energy.

The five-dimensional fission potential-energy surface at and beyond the several outer saddle points is an imposing landscape traversed by deep valleys and high ridges. The fragment shells profoundly influence the topography of this landscape, long before the system divides into separated fission fragments.

Note added in proof: An interesting discussion of the problem of finding structures in multidimensional spaces is in ref. 37. □

Received 3 August; accepted 18 December 2000.

1. Meitner, L. & Frisch, O. R. Disintegration of uranium by neutrons: A new type of nuclear reaction. *Nature* **143**, 239–240 (1939).
2. Bohr, N. & Wheeler, J. A. The mechanism of fission. *Phys. Rev.* **56**, 426–450 (1939).
3. Hahn, O. & Strassmann, F. Über den Nachweis und das Verhalten der bei der Bestrahlung des Urans mittels Neutronen entstehenden Erdalkalimetalle. *Naturwissenschaften* **27**, 11–15 (1939).
4. Frankel, S. & Metropolis, N. Liquid-drop model of fission. *Phys. Rev.* **72**, 914–925 (1947).
5. Strutinsky, V. M. Shell effects in nuclear masses and deformation energies. *Nucl. Phys. A* **95**, 420–442 (1967).
6. Strutinsky, V. M. Shells in deformed nuclei. *Nucl. Phys. A* **122**, 1–33 (1968).
7. Nilsson, S. G. *et al.* On the nuclear structure and stability of heavy and superheavy elements. *Nucl. Phys. A* **131**, 1–66 (1969).

8. Pashkevich, V. V. The energy of non-axial deformation of heavy nuclei. *Nucl. Phys. A* **133**, 400–404 (1969).
9. Möller, P. & Nilsson, S. G. The fission barrier and odd-multipole shape distortions. *Phys. Lett.* **31B**, 283–286 (1970).
10. Brack, M. *et al.* Funny hills: The shell-correction approach to nuclear shell effects and its applications to the fission process. *Rev. Mod. Phys.* **44**, 320–405 (1972).
11. Nix, J. R. Calculation of fission barriers for heavy and superheavy nuclei. *Annu. Rev. Nucl. Sci.* **22**, 65–120 (1972).
12. Möller, P., Nix, J. R., Myers, W. D. & Swiatecki, W. J. Nuclear ground-state masses and deformations. *Atom. Data Nucl. Data Tables* **59**, 185–381 (1995).
13. Aboussir, Y., Pearson, J. M., Dutta, A. K. & Tondeur, F. Nuclear-mass formula via an approximation to the Hartree-Fock method. *Atom. Data Nucl. Data Tables* **61**, 127–176 (1995).
14. Möller, P. & Nix, J. R. Potential-energy surfaces for asymmetric heavy-ion reactions. *Nucl. Phys. A* **281**, 354–372 (1977).
15. Möller, P., Nix, J. R. & Swiatecki, W. J. Calculated fission properties of the heaviest elements. *Nucl. Phys. A* **469**, 1–50 (1987).
16. Möller, P., Nix, J. R. & Swiatecki, W. J. New developments in the calculation of heavy-element fission barriers. *Nucl. Phys. A* **492**, 349–387 (1989).
17. Armbruster, P. Nuclear structure in cold rearrangement processes in fission and fusion. *Rep. Prog. Phys.* **62**, 465–525 (1999).
18. Britt, H. C., Wegner, H. E. & Gursky, J. C. Energetics of charged particle-induced fission reactions. *Phys. Rev.* **129**, 2239–2252 (1963).
19. Konecny, E., Specht, H. J. & Weber, J. in *Proc. Third IAEA Symp. Phys. Chem. Fission* Vol. II, 3–18 (International Atomic Energy Agency, Vienna, 1974).
20. Ohtsuki, T., Nakahara, H. & Nagame, Y. Systematic variation of fission barrier heights for symmetrical and asymmetric mass divisions. *Phys. Rev. C* **48**, 1667–1676 (1993).
21. Nagame, Y. *et al.* Bimodal nature of low energy fission of light actinides. *Radiochim. Acta* **78**, 3–10 (1997).
22. Möller, P. & Iwamoto, A. Realistic fission saddle-point shapes. *Phys. Rev. C* **61**, 47602-1–4 (2000).
23. Möller, P., Nix, J. R. & Kratz, K.-L. Nuclear properties for astrophysical and radioactive-ion-beam applications. *Atom. Data Nucl. Data Tables* **66**, 131–343 (1997).
24. Mamdouh, A., Pearson, J. M., Rayet, M. & Tondeur, F. Large-scale fission-barrier calculations with the ETFSI method. *Nucl. Phys. A* **644**, 389–414 (1998).
25. Hulet, E. K. *et al.* Bimodal symmetrical fission observed in the heaviest elements. *Phys. Rev. Lett.* **56**, 313–316 (1986).
26. Hoffman, D. C. & Hoffman, M. M. Post-fission phenomena. *Annu. Rev. Nucl. Sci.* **24**, 151–207 (1974).
27. Dematte, L., Wagemans, C., Barthelemy, R., Dhondt, P. & Deruyter, A. Fragments' mass and energy characteristics in the spontaneous fission of Pu-236, Pu-238, Pu-240, Pu-242 and Pu-244. *Nucl. Phys. A* **617**, 331–346 (1997).
28. Schmidt, K.-H. *et al.* Relativistic radioactive beams: A new access to nuclear-fission studies. *Nucl. Phys. A* **665**, 221–267 (2000).
29. Bolsterli, M., Fiset, E. O., Nix, J. R. & Norton, J. L. New calculation of fission barriers for heavy and superheavy nuclei. *Phys. Rev. C* **5**, 1050–1075 (1972).
30. Krappe, H. J., Nix, J. R. & Sierk, A. J. Unified nuclear potential for heavy-ion elastic scattering, fusion, fission, and ground-state masses and deformations. *Phys. Rev. C* **20**, 992–1013 (1979).
31. Möller, P. & Iwamoto, A. Macroscopic potential-energy surfaces for arbitrarily oriented, deformed heavy-ions. *Nucl. Phys. A* **575**, 381–411 (1994).
32. Möller, P. & Randrup, J. New developments in the calculation of β -strength functions. *Nucl. Phys. A* **514**, 1–48 (1990).
33. Kratz, K.-L., Bitouzet, J.-P., Thielemann, F.-K., Möller, P. & Pfeiffer, B. Isotopic r-process abundances and nuclear-structure far from stability: implications for the r-process mechanism. *Astrophys. J.* **403**, 216–238 (1993).
34. Kadyaev, G. A., Ostapenko, Yu. B. & Smirenkin, G. N. Thresholds and saddle shapes in symmetric and asymmetric fission in the vicinity of Ra. *Sov. J. Nucl. Phys.* **45**, 951–958 (1987).
35. Zhao, Y. L. *et al.* Experimental verification of two deformation paths in the mass division process of actinides. *J. Alloys Comp.* **271**, 327–330 (1998).
36. Meitner, L. Fission and nuclear shell model. *Nature* **165**, 561 (1950).
37. Hayes, B. Dividing the continent. *Am. Sci.* **88**, 481–485 (2000).

Acknowledgements

The calculations on which the results in this paper are based were carried out on the cluster of 4 Alpha processors at the TANDEM accelerator in JAERI in the winter of 1998–99 and subsequently on the AVALON cluster of 140 Alpha processors at Los Alamos. This research is supported by the US DOE.

Correspondence and requests for materials should be addressed to P.M. (e-mail: molle@moller.lanl.gov).

Experimental violation of a Bell's inequality with efficient detection

M. A. Rowe*, D. Kielpinski*, V. Meyer*, C. A. Sackett*, W. M. Itano*, C. Monroe† & D. J. Wineland*

* Time and Frequency Division, National Institute of Standards and Technology, Boulder, Colorado 80305, USA

† Department of Physics, University of Michigan, Ann Arbor, Michigan 48109, USA

Local realism is the idea that objects have definite properties whether or not they are measured, and that measurements of these properties are not affected by events taking place sufficiently far away¹. Einstein, Podolsky and Rosen² used these reasonable assumptions to conclude that quantum mechanics is incomplete. Starting in 1965, Bell and others constructed mathematical inequalities whereby experimental tests could distinguish between quantum mechanics and local realistic theories^{1,3–5}. Many experiments^{1,6–15} have since been done that are consistent with quantum mechanics and inconsistent with local realism. But these conclusions remain the subject of considerable interest and debate, and experiments are still being refined to overcome ‘loopholes’ that might allow a local realistic interpretation. Here we have measured correlations in the classical properties of massive entangled particles (⁹Be⁺ ions): these correlations violate a form of Bell's inequality. Our measured value of the appropriate Bell's ‘signal’ is 2.25 ± 0.03 , whereas a value of 2 is the maximum allowed by local realistic theories of nature. In contrast to previous measurements with massive particles, this violation of Bell's inequality was obtained by use of a complete set of measurements. Moreover, the high detection efficiency of our apparatus eliminates the so-called ‘detection’ loophole.

Early experiments to test Bell's inequalities were subject to two primary, although seemingly implausible, loopholes. The first might be termed the locality or ‘lightcone’ loophole, in which the correlations of apparently separate events could result from unknown subluminal signals propagating between different regions of the apparatus. Aspect¹⁶ has given a brief history of this issue, starting with the experiments of ref. 8 and highlighting the strict relativistic separation between measurements reported by the Innsbruck group¹⁵. Similar results have also been reported for the Geneva experiment^{14,17}. The second loophole is usually referred to as the detection loophole. All experiments up to now have had detection efficiencies low enough to allow the possibility that the subensemble of detected events agrees with quantum mechanics even though the entire ensemble satisfies Bell's inequalities. Therefore it must be assumed that the detected events represent the entire ensemble; a fair-sampling hypothesis. Several proposals for closing this loophole have been made^{18–24}; we believe the experiment that we report here is the first to do so. Another feature of our experiment is that it uses massive particles. A previous test of Bell's inequality was carried out on protons²⁵, but the interpretation of the detected events relied on quantum mechanics, as symmetries valid given quantum mechanics were used to extrapolate the data to a complete set of Bell's angles. Here we do not make such assumptions.

A Bell measurement of the type suggested by Clauser, Horne, Shimony and Holt⁵ (CHSH) consists of three basic ingredients (Fig. 1a). First is the preparation of a pair of particles in a repeatable starting configuration (the output of the ‘magic’ box in Fig. 1a). Second, a variable classical manipulation is applied independently to each particle; these manipulations are labelled ϕ_1 and ϕ_2 . Finally, in the detection phase, a classical property with two possible outcomes is measured for each of the particles. The correlation of these outcomes

$$q(\phi_1, \phi_2) = \frac{N_{\text{same}}(\phi_1, \phi_2) - N_{\text{different}}(\phi_1, \phi_2)}{N_{\text{same}} + N_{\text{different}}} \quad (1)$$

is measured by repeating the experiment many times. Here N_{same} and $N_{\text{different}}$ are the number of measurements where the two results were the same and different, respectively. The CHSH form of Bell's inequalities states that the correlations resulting from local realistic theories must obey:

$$B(\alpha_1, \delta_1, \beta_2, \gamma_2) = |q(\delta_1, \gamma_2) - q(\alpha_1, \gamma_2)| + |q(\delta_1, \beta_2) + q(\alpha_1, \beta_2)| \leq 2 \quad (2)$$

where α_1 and δ_1 (β_2 and γ_2) are specific values of ϕ_1 (ϕ_2). For example, in a photon experiment¹⁵, parametric down-conversion prepares a pair of photons in a singlet Einstein–Podolsky–Rosen (EPR) pair. After this, a variable rotation of the photon polarization is applied to each photon. Finally, the photons' polarization states, vertical or horizontal, are determined.

Our experiment prepares a pair of two-level atomic ions in a repeatable configuration (entangled state). Next, a laser field is applied to the particles; the classical manipulation variables are the

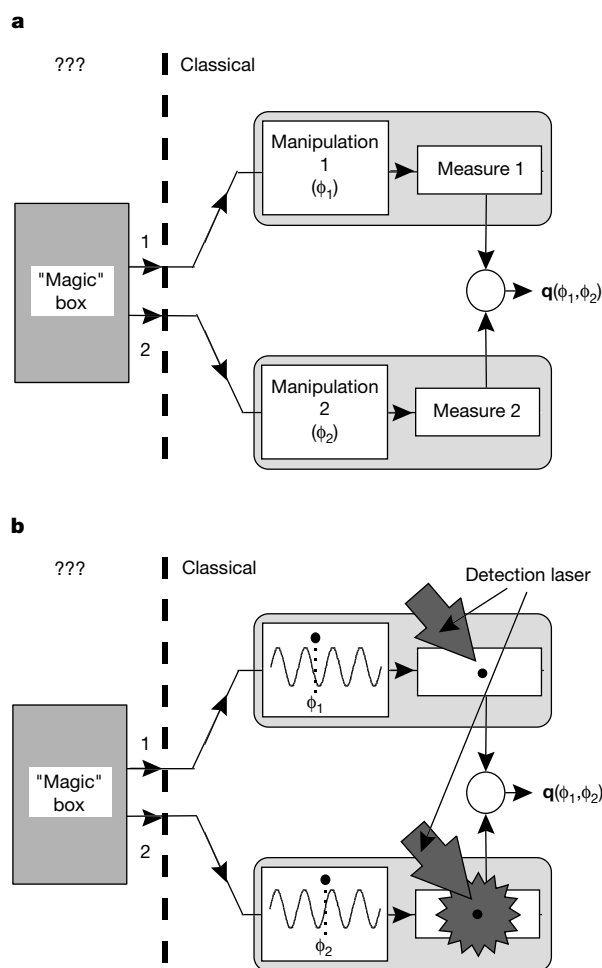


Figure 1 Illustration of how Bell's inequality experiments work. The idea is that a ‘magic box’ emits a pair of particles. We attempt to determine the joint properties of these particles by applying various classical manipulations to them and observing the correlations of the measurement outputs. **a**, A general CHSH type of Bell's inequality experiment. **b**, Our experiment. The manipulation is a laser wave applied with phases ϕ_1 and ϕ_2 to ion 1 and ion 2 respectively. The measurement is the detection of photons emanating from the ions upon application of a detection laser. Two possible measurement outcomes are possible, detection of few photons (as depicted for ion 1 in the figure) or the detection of many photons (as depicted for ion 2 in the figure).

phases of this field at each ion's position. Finally, upon application of a detection laser beam, the classical property measured is the number of scattered photons emanating from the particles (which effectively measures their atomic states). Figure 1b shows how our experiment maps onto the general case. Entangled atoms produced in the context of cavity-quantum-electrodynamics²⁶ could similarly be used to measure Bell's inequalities.

The experimental apparatus is as described in ref. 27. Two $^9\text{Be}^+$ ions are confined along the axis of a linear Paul trap with an axial centre-of-mass frequency of 5 MHz. We select two resolved levels of the $2S_{1/2}$ ground state, $|\downarrow\rangle \equiv |F=2, m_F=-2\rangle$ and $|\uparrow\rangle \equiv |F=1, m_F=-1\rangle$, where F and m_F are the quantum numbers of the total angular momentum. These states are coupled by a coherent stimulated Raman transition. The two laser beams used to drive the transition have a wavelength of 313 nm and a difference frequency near the hyperfine splitting of the states, $\omega_0 \approx 2\pi \times 1.25$ GHz. The beams are aligned perpendicular to each other, with their difference wavevector $\Delta\mathbf{k}$ along the trap axis. As described in ref. 27, it is possible in this configuration to produce the entangled state

$$|\psi_2\rangle = \frac{1}{\sqrt{2}}(|\uparrow\uparrow\rangle - |\downarrow\downarrow\rangle) \quad (3)$$

The fidelity $F = \langle\psi_2|\rho|\psi_2\rangle$, where ρ is the density matrix for the state we make, was about 88% for the data runs. In the discussion below we assume $|\psi_2\rangle$ as the starting condition for the experiment.

After making the state $|\psi_2\rangle$, we again apply Raman beams for a pulse of short duration (~ 400 ns) so that the state of each ion j is

transformed in the interaction picture as

$$|\uparrow_j\rangle \rightarrow \frac{1}{\sqrt{2}}(|\uparrow_j\rangle - ie^{-i\phi_j}|\downarrow_j\rangle); |\downarrow_j\rangle \rightarrow \frac{1}{\sqrt{2}}(|\downarrow_j\rangle - ie^{i\phi_j}|\uparrow_j\rangle) \quad (4)$$

The phase, ϕ_j , is the phase of the field driving the Raman transitions (more specifically, the phase difference between the two Raman beams) at the position of ion j and corresponds to the inputs ϕ_1 and ϕ_2 in Fig. 1. We set this phase in two ways in the experiment. First, as an ion is moved along the trap axis this phase changes by $\Delta\mathbf{k}\cdot\Delta\mathbf{x}_j$. For example, a translation of $\lambda/\sqrt{2}$ along the trap axis corresponds to a phase shift of 2π . In addition, the laser phase on both ions is changed by a common amount by varying the phase, ϕ_s , of the radio-frequency synthesizer that determines the Raman difference frequency. The phase on ion j is therefore

$$\phi_j = \phi_s + \Delta\mathbf{k}\cdot\mathbf{x}_j \quad (5)$$

In the experiment, the axial trap strength is changed so that the ions move about the centre of the trap symmetrically, giving $\Delta\mathbf{x}_1 = -\Delta\mathbf{x}_2$. Therefore the trap strength controls the differential phase, $\Delta\phi \equiv \phi_1 - \phi_2 = \Delta\mathbf{k}\cdot(\mathbf{x}_1 - \mathbf{x}_2)$, and the synthesizer controls the total phase, $\phi_{\text{tot}} \equiv \phi_1 + \phi_2 = 2\phi_s$. The calibration of these relations is discussed in the Methods.

The state of an ion, $|\downarrow\rangle$ or $|\uparrow\rangle$, is determined by probing the ion with circularly polarized light from a 'detection' laser beam²⁷. During this detection pulse, ions in the $|\downarrow\rangle$ or bright state scatter many photons, and on average about 64 of these are detected with a photomultiplier tube, while ions in the $|\uparrow\rangle$ or dark state scatter very few photons. For two ions, three cases can occur: zero ions bright, one ion bright, or two ions bright. In the one-ion-bright case it is not necessary to know which ion is bright because the Bell's measurement requires only knowledge of whether or not the ions' states are different. Figure 2 shows histograms, each with 20,000 detection measurements. The three cases are distinguished from each other with simple discriminator levels in the number of photons collected with the phototube.

An alternative description of our experiment can be made in the language of spin-one-half magnetic moments in a magnetic field (directed in the $\hat{z}$ direction). The dynamics of the spin system are the same as for our two-level system²⁸. Combining the manipulation (equation (4)) and measurement steps, we effectively measure the spin projection of each ion j in the $\hat{r}_j$ direction, where the vector $\hat{r}_j$ is in the $\hat{x}-\hat{y}$ plane at an angle ϕ_j to the $\hat{y}$ axis. Although we have used quantum-mechanical language to describe the manipulation and measurement steps, we emphasize that both are procedures completely analogous to the classical rotations of wave-plates and measurements of polarization in an optical apparatus.

Here we calculate the quantum-mechanical prediction for the correlation function. Our manipulation step transforms the starting

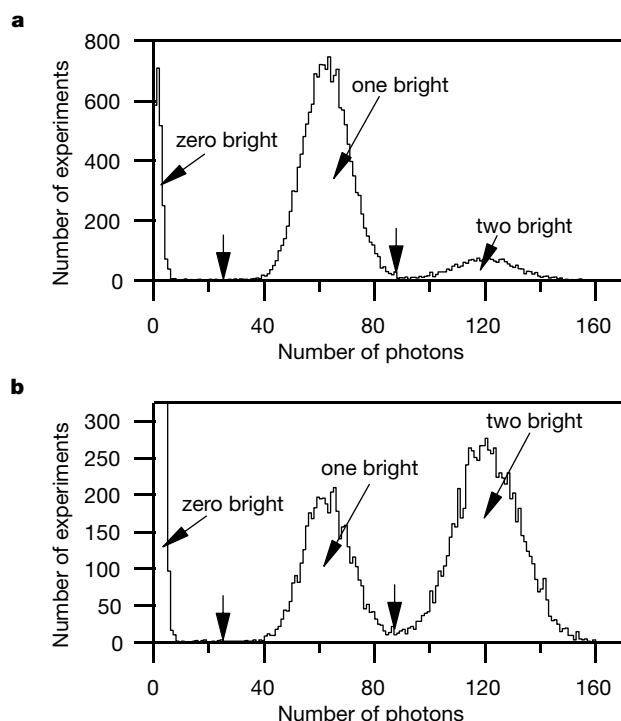


Figure 2 Typical data histograms comprising the detection measurements of 20,000 experiments taking a total time of about 20 s. In each experiment the population in the $|\uparrow\rangle$ state is first coherently transferred to the $|F=1, M_F=+1\rangle$ to make it even less likely to fluoresce upon application of the detection laser. The detection laser is turned on and the number of fluorescence photons detected by the phototube in 1 ms is recorded. The cut between the one bright and two bright cases is made so that the fractions of two equal distributions which extend past the cut points are equal. The vertical arrows indicate the location of the cut between the 0 (1) bright and 1 (2) bright peaks at 25 (86) counts.

a, Data histogram with a negative correlation using $\phi_1 = 3\pi/8$ and $\phi_2 = 3\pi/8$. For these data $N_0 \approx 2,200$, $N_1 \approx 15,500$ and $N_2 \approx 2,300$. **b**, Data histogram with a positive correlation using $\phi_1 = 3\pi/8$ and $\phi_2 = -\pi/8$. For these data $N_0 \approx 7,700$, $N_1 \approx 4,400$ and $N_2 \approx 7,900$. The zero bright peak extends vertically to 2,551.

Table 1 The four sets of phase angles used for the Bell's experiment

Experiment input	ϕ_1	ϕ_2	$\Delta\phi$	ϕ_{tot}
$\alpha_1\beta_2$	$-\pi/8$	$-\pi/8$	0	$-\pi/4$
$\alpha_1\gamma_2$	$-\pi/8$	$3\pi/8$	$-\pi/2$	$+\pi/4$
$\delta_1\beta_2$	$3\pi/8$	$-\pi/8$	$+\pi/2$	$+\pi/4$
$\delta_1\gamma_2$	$3\pi/8$	$3\pi/8$	0	$+3\pi/4$

Table 2 Correlation values and resulting Bell's signals for five experimental runs

Run number	$q(\alpha_1, \beta_2)$	$q(\alpha_1, \gamma_2)$	$q(\delta_1, \beta_2)$	$q(\delta_1, \gamma_2)$	$B(\alpha_1, \delta_1, \beta_2, \gamma_2)$
1	0.541	0.539	0.569	-0.573	2.222
2	0.575	0.570	0.530	-0.600	2.275
3	0.551	0.634	0.590	-0.487	2.262
4	0.575	0.561	0.559	-0.551	2.246
5	0.541	0.596	0.537	-0.571	2.245

The experimental angle values were $\alpha_1 = -(\pi/8)$, $\delta_1 = 3\pi/8$, $\beta_2 = -(\pi/8)$, and $\gamma_2 = 3\pi/8$. The statistical errors are 0.006 and 0.012 for the q and B values respectively. The systematic errors (see text) are 0.03 and 0.06 for the q and B values respectively.

state, $|\psi_2\rangle$, to

$$|\psi_2'\rangle = \frac{1}{2\sqrt{2}} \{ (1 + e^{i(\phi_1+\phi_2)}) (|\uparrow\uparrow\rangle - e^{-i(\phi_1+\phi_2)} |\downarrow\downarrow\rangle) - i(1 - e^{i(\phi_1+\phi_2)}) (e^{-i\phi_2} |\uparrow\downarrow\rangle + e^{-i\phi_1} |\downarrow\uparrow\rangle) \} \quad (6)$$

Using the measurement operators $\hat{N}_{\text{same}} = N_{\text{tot}}[|\uparrow\uparrow\rangle\langle\uparrow\uparrow| + |\downarrow\downarrow\rangle\langle\downarrow\downarrow|]$ and $\hat{N}_{\text{different}} = N_{\text{tot}}[|\uparrow\downarrow\rangle\langle\uparrow\downarrow| + |\downarrow\uparrow\rangle\langle\downarrow\uparrow|]$, the correlation function is calculated to be

$$q(\phi_1, \phi_2) = \frac{1}{8} [2|1 + e^{i(\phi_1+\phi_2)}|^2 - 2|1 - e^{i(\phi_1+\phi_2)}|^2] = \cos(\phi_1 + \phi_2) \quad (7)$$

The CHSH inequality (equation (2)) is maximally violated by quantum mechanics at certain sets of phase angles. One such set is $\alpha_1 = -(\pi/8)$, $\delta_1 = 3\pi/8$, $\beta_2 = -(\pi/8)$ and $\gamma_2 = 3\pi/8$. With these phase angles quantum mechanics predicts

$$B\left(-\frac{\pi}{8}, \frac{3\pi}{8}, -\frac{\pi}{8}, \frac{3\pi}{8}\right) = 2\sqrt{2} \quad (8)$$

This violates the local realism condition, which requires that $B \leq 2$.

The correlation function is measured experimentally at four sets of phase angles, listed in Table 1. The experiment is repeated $N_{\text{tot}} = 20,000$ times at each of the four sets of phases. For each set of phases the correlation function is calculated using

$$q = \frac{(N_0 + N_2) - N_1}{N_{\text{tot}}} \quad (9)$$

Here N_0 , N_1 and N_2 are the number of events with zero, one and two ions bright, respectively. The correlation values from the four sets of phase angles are combined into the Bell's signal, $B(\alpha_1, \delta_1, \beta_2, \gamma_2)$, using equation (2). The correlation values and resulting Bell's signals from five data runs are given in Table 2.

So far we have described the experiment in terms of perfect implementation of the phase angles. In the actual experiment, however, α_1 , δ_1 , β_2 and γ_2 are not quite the same angles both times they occur in the Bell's inequality. In our experiment the dominant reason for this error results from the phase instability of the synthesizer, which can cause the angles to drift appreciably during four minutes, the time required to take a complete set of measurements. This random drift causes a root-mean-squared error for the correlation function of ± 0.03 on this timescale, which propagates to an error of ± 0.06 for the Bell's signal. The error for the Bell's signal from the five combined data sets is then ± 0.03 , consistent with the run-to-run variation observed. Averaging the five Bell's signals from Table 2, we arrive at our experimental result, which is

$$B\left(-\frac{\pi}{8}, \frac{3\pi}{8}, -\frac{\pi}{8}, \frac{3\pi}{8}\right) = 2.25 \pm 0.03 \quad (10)$$

If we take into account the imperfections of our experiment (imperfect state fidelity, manipulations, and detection), this value agrees with the prediction of quantum mechanics.

The result above was obtained using the outcomes of every experiment, so that no fair-sampling hypothesis is required. In this case, the issue of detection efficiency is replaced by detection accuracy. The dominant cause of inaccuracy in our state detection comes from the bright state becoming dark because of optical pumping effects. For example, imperfect circular polarization of the detection light allows an ion in the $|\downarrow\rangle$ state to be pumped to $|\uparrow\rangle$, resulting in fewer collected photons from a bright ion. Because of such errors, a bright ion is misidentified 2% of the time as being dark. This imperfect detection accuracy decreases the magnitude of the measured correlations. We estimate that our Bell's signal would be 2.37 with perfect detection accuracy.

We have thus presented experimental results of a Bell's inequality

measurement where a measurement outcome was recorded for every experiment. Our detection efficiency was high enough for a Bell's inequality to be violated without requiring the assumption of fair sampling, thereby closing the detection loophole in this experiment. The ions were separated by a distance large enough that no known interaction could affect the results; however, the lightcone loophole remains open here. Further details of this experiment will be published elsewhere. $\square$

Methods

Phase calibration

The experiment was run with specific phase differences of the Raman laser beam fields at each ion. In order to implement a complete set of laser phases, a calibration of the phase on each ion as a function of axial trap strength was made. We emphasize that the calibration method is classical in nature. Although quantum mechanics guided the choice of calibration method, no quantum mechanics was used to interpret the signal. General arguments are used to describe the signal resulting from a sequence of laser pulses and its dependence on the classical physical parameters of the system, the laser phase at the ion, and the ion's position.

In the calibration procedure, a Ramsey experiment was performed on two ions. The first $\pi/2$ Rabi rotation was performed identically each time. The laser phases at the ions' positions for the second $\pi/2$ Rabi rotation were varied, ϕ_1 for ion 1 and ϕ_2 for ion 2. The detection signal is the total number of photons counted during detection. With an auxiliary one-ion experiment we first established empirically that the individual signal depends only on the laser phase at an individual ion and is $C + A\cos\phi_i$. Here C and A are the offset and amplitude of the one-ion signal. We measure the detector to be linear, so that the detection signal is the sum of the two ions' individual signals. The two-ion signal is therefore

$$C + A\cos\phi_1 + C + A\cos\phi_2 = 2C + 2A\cos\left[\frac{1}{2}(\phi_1 + \phi_2)\right]\cos\left[\frac{1}{2}(\phi_1 - \phi_2)\right] \quad (11)$$

By measuring the fringe amplitude and phase as $\phi = (\phi_1 + \phi_2)/2$ is swept, we calibrate $\phi_1 - \phi_2$ as a function of trap strength and ensure that $\phi_1 + \phi_2$ is independent of trap strength.

We use the phase convention that at the ion separation used for the entanglement preparation pulse the maximum of the correlation function is at $\phi_1 = \phi_2 = 0$ (or $\Delta\phi = \phi_{\text{tot}} = 0$). Our measurement procedure begins by experimentally finding this condition of $\phi_1 = \phi_2 = 0$ by keeping $\Delta\phi = 0$ and scanning the synthesizer phase to find the maximum correlation. The experiment is then adjusted to the phase angles specified above by switching the axial trap strength to set $\Delta\phi$ and incrementing the synthesizer phase to set ϕ_{tot} .

Locality issues

The ions are separated by a distance of approximately $3\text{ }\mu\text{m}$, which is greater than 100 times the size of the wavepacket of each ion. Although the Coulomb interaction strongly couples the ions' motion it does not affect the ions' internal states. At this distance, all known relevant interactions are expected to be small. For example, dipole-dipole interactions between the ions slightly modify the light-scattering intensity, but this effect is negligible for the ion-ion separations used²⁹. Also, the detection solid angle is large enough that Young's interference fringes, if present, are averaged out³⁰. Even though all known interactions would cause negligible correlations in the measurement outcomes, the ion separation is not large enough to eliminate the lightcone loophole.

We note that the experiment would be conceptually simpler if, after creating the entangled state, we separated the ions so that the input manipulations and measurements were done individually. However, unless we separated the ions by a distance large enough to overcome the lightcone loophole, this is only a matter of convenience of description and does not change the conclusions that can be drawn from the results.

Received 25 October; accepted 30 November 2000.

1. Clauser, J. F. & Shimony, A. Bell's theorem: experimental tests and implications. *Rep. Prog. Phys.* **41**, 1883–1927 (1978).
2. Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of reality be considered complete? *Phys. Rev.* **47**, 777–780 (1935).
3. Bell, J. S. On the Einstein-Podolsky-Rosen paradox. *Physics* **1**, 195–200 (1965).
4. Bell, J. S. in *Foundations of Quantum Mechanics* (ed. d'Espagnat, B.) 171–181 (Academic, New York, 1971).
5. Clauser, J. F., Horne, M. A., Shimony, A. & Holt, R. A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23**, 880–884 (1969).
6. Freedman, S. J. & Clauser, J. F. Experimental test of local hidden-variable theories. *Phys. Rev. Lett.* **28**, 938–941 (1972).
7. Fry, E. S. & Thompson, R. C. Experimental test of local hidden-variable theories. *Phys. Rev. Lett.* **37**, 465–468 (1976).
8. Aspect, A., Grangier, P. & Roger, G. Experimental realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: a new violation of Bell's inequalities. *Phys. Rev. Lett.* **49**, 91–94 (1982).
9. Aspect, A., Dalibard, J. & Roger, G. Experimental test of Bell's inequalities using time-varying analyzers. *Phys. Rev. Lett.* **49**, 1804–1807 (1982).
10. Ou, Z. Y. & Mandel, L. Violation of Bell's inequality and classical probability in a two-photon correlation experiment. *Phys. Rev. Lett.* **61**, 50–53 (1988).
11. Shih, Y. H. & Alley, C. O. New type of Einstein-Podolsky-Rosen-Bohm experiment using pairs of light

- quanta produced by optical parametric down conversion. *Phys. Rev. Lett.* **61**, 2921–2924 (1988).
12. Tapster, P. R., Rarity, J. G. & Owens, P. C. M. Violation of Bell's inequality over 4 km of optical fiber. *Phys. Rev. Lett.* **73**, 1923–1926 (1994).
 13. Kwiat, P. G., Mattle, K., Weinfurter, H. & Zeilinger, A. New high-intensity source of polarization-entangled photon pairs. *Phys. Rev. Lett.* **75**, 4337–4341 (1995).
 14. Tittel, W., Brendel, J., Zbinden, H. & Gisin, N. Violation of Bell inequalities by photons more than 10 km apart. *Phys. Rev. Lett.* **81**, 3563–3566 (1998).
 15. Weihs, G. *et al.* Violation of Bell's inequality under strict Einstein locality conditions. *Phys. Rev. Lett.* **81**, 5039–5043 (1998).
 16. Aspect, A. Bell's inequality test: more ideal than ever. *Nature* **398**, 189–190 (1999).
 17. Gisin, N. & Zbinden, H. Bell inequality and the locality loophole: active versus passive switches. *Phys. Lett. A* **264**, 103–107 (1999).
 18. Lo, T. K. & Shimony, A. Proposed molecular test of local hidden-variable theories. *Phys. Rev. A* **23**, 3003–3012 (1981).
 19. Kwiat, P. G., Eberhard, P. H., Steinberg, A. M. & Chiao, R. Y. Proposal for a loophole-free Bell inequality experiment. *Phys. Rev. A* **49**, 3209–3220 (1994).
 20. Huelga, S. F., Ferrero, M. & Santos, E. Loophole-free test of the Bell inequality. *Phys. Rev. A* **51**, 5008–5011 (1995).
 21. Fry, E. S., Walther, T. & Li, S. Proposal for a loophole free test of the Bell inequalities. *Phys. Rev. A* **52**, 4381–4395 (1995).
 22. Freyberger, M., Aravind, P. K., Horne, M. A. & Shimony, A. Proposed test of Bell's inequality without a detection loophole by using entangled Rydberg atoms. *Phys. Rev. A* **53**, 1232–1244 (1996).
 23. Brif, C. & Mann, A. Testing Bell's inequality with two-level atoms via population spectroscopy. *Europhys. Lett.* **49**, 1–7 (2000).
 24. Beige, A., Munro, W. J. & Knight, P. L. A Bell's inequality test with entangled atoms. *Phys. Rev. A* **62**, 052102-1–052102-9 (2000).
 25. Lamehi-Rachti, M. & Mittag, W. Quantum mechanics and hidden variables: a test of Bell's inequality by the measurement of the spin correlation in low-energy proton–proton scattering. *Phys. Rev. D* **14**, 2543–2555 (1976).
 26. Hagley, E. *et al.* Generation of Einstein-Podolsky-Rosen pairs of atoms. *Phys. Rev. Lett.* **79**, 1–5 (1997).
 27. Sackett, C. A. *et al.* Experimental entanglement of four particles. *Nature* **404**, 256–259 (2000).
 28. Feynman, R. P., Vernon, F. L. & Hellwarth, R. W. Geometrical representation of the Schrödinger equation for solving maser problems. *J. Appl. Phys.* **28**, 49–52 (1957).
 29. Richter, T. Cooperative resonance fluorescence from two atoms experiencing different driving fields. *Optica Acta* **30**, 1769–1780 (1983).
 30. Eichmann, U. *et al.* Young's interference experiment with light scattered from two atoms. *Phys. Rev. Lett.* **70**, 2359–2362 (1993).

Acknowledgements

We thank A. Ben-Kish, J. Bollinger, J. Britton, N. Gisin, P. Knight, P. Kwiat and I. Percival for useful discussions and comments on the manuscript. This work was supported by the US National Security Agency (NSA) and the Advanced Research and Development Activity (ARDA), the US Office of Naval Research, and the US Army Research Office. This paper is a contribution of the National Institute of Standards and Technology and is not subject to US copyright.

Correspondence and requests for materials should be addressed to D.J.W. (e-mail: david.wineland@boulder.nist.gov).

Autonomic healing of polymer composites

S. R. White*, N. R. Sottos†, P. H. Geubelle*, J. S. Moore‡, M. R. Kessler†, S. R. Sriram‡, E. N. Brown† & S. Viswanathan*

* Department of Aeronautical and Astronautical Engineering, † Department of Theoretical and Applied Mechanics, ‡ Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA

Structural polymers are susceptible to damage in the form of cracks, which form deep within the structure where detection is difficult and repair is almost impossible. Cracking leads to mechanical degradation^{1–3} of fibre-reinforced polymer composites; in microelectronic polymeric components it can also lead to electrical failure⁴. Microcracking induced by thermal and mechanical fatigue is also a long-standing problem in polymer adhesives⁵. Regardless of the application, once cracks have formed within polymeric materials, the integrity of the structure is significantly compromised. Experiments exploring the concept of self-repair have been previously reported^{6–8}, but the only successful crack-healing methods that have been reported so far

require some form of manual intervention^{10–18}. Here we report a structural polymeric material with the ability to autonomically heal cracks. The material incorporates a microencapsulated healing agent that is released upon crack intrusion. Polymerization of the healing agent is then triggered by contact with an embedded catalyst, bonding the crack faces. Our fracture experiments yield as much as 75% recovery in toughness, and we expect that our approach will be applicable to other brittle materials systems (including ceramics and glasses).

Figure 1 illustrates our autonomic healing concept. Healing is accomplished by incorporating a microencapsulated healing agent and a catalytic chemical trigger within an epoxy matrix. An approaching crack ruptures embedded microcapsules, releasing healing agent into the crack plane through capillary action. Polymerization of the healing agent is triggered by contact with the embedded catalyst, bonding the crack faces. The damage-induced triggering mechanism provides site-specific autonomic control of repair. An additional unique feature of our healing concept is the use of living (that is, having unterminated chain-ends) polymerization catalysts, thus enabling multiple healing events. Engineering this self-healing composite involves the challenge of combining polymer science, experimental and analytical mechanics, and composites processing principles.

We began by analysing the effects of microcapsule geometry and properties on the mechanical triggering process. For example, capsule walls that are too thick will not rupture when the crack approaches, whereas capsules with very thin walls will break during processing. Other relevant design parameters are the toughness and the relative stiffness of the microcapsules, and the strength of the interface between the microcapsule and the matrix. Micro-mechanical modelling with the aid of the Eshelby–Mura equivalent inclusion method¹⁹ has been used to study various aspects of the complex three-dimensional interaction between a crack and a microcapsule. An illustrative result from these studies is presented in Fig. 2a, which shows the effect of the relative stiffness of the microcapsule on the propagation path of an approaching crack. The crack, the sphere and the surrounding matrix are subjected to a far-field tensile loading, σ_∞ , perpendicular to the crack plane.

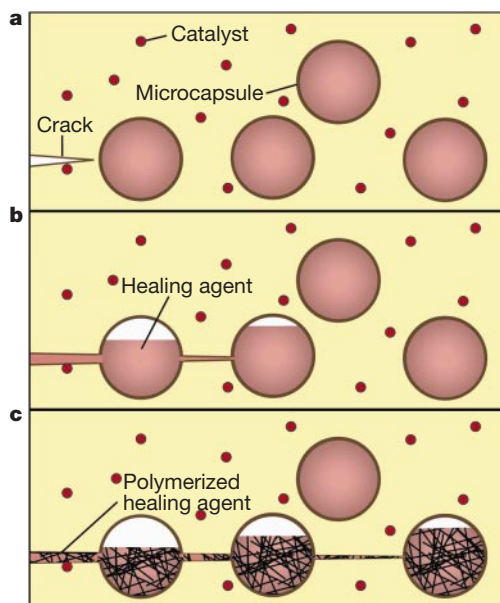


Figure 1 The autonomic healing concept. A microencapsulated healing agent is embedded in a structural composite matrix containing a catalyst capable of polymerizing the healing agent. **a**, Cracks form in the matrix wherever damage occurs; **b**, the crack ruptures the microcapsules, releasing the healing agent into the crack plane through capillary action; **c**, the healing agent contacts the catalyst, triggering polymerization that bonds the crack faces closed.

As apparent from the σ_{22} stress distribution in the equatorial plane of the sphere in Fig. 2a, the stiffness of the sphere relative to the matrix strongly affects the stress state in the proximity of the crack tip and around the sphere itself. In the case of a stiffer inclusion, the stress field in the immediate vicinity of the crack tip shows an asymmetry that indicates an undesirable tendency of the crack to be deflected away from the inclusion. The situation is reversed in the case of a more compliant spherical inclusion, and the crack is

attracted toward the microcapsule, a necessary condition for its rupture and the triggering of the healing process.

Observations by optical and scanning electron microscopy confirmed the healing concept in Fig. 1 and substantiated the main findings from analytical studies in Fig. 2a. The time sequence of optical images in Fig. 2b shows the rupture of an embedded microcapsule filled with a red dye and the subsequent release of the healing agent into the crack plane. The scanning electron

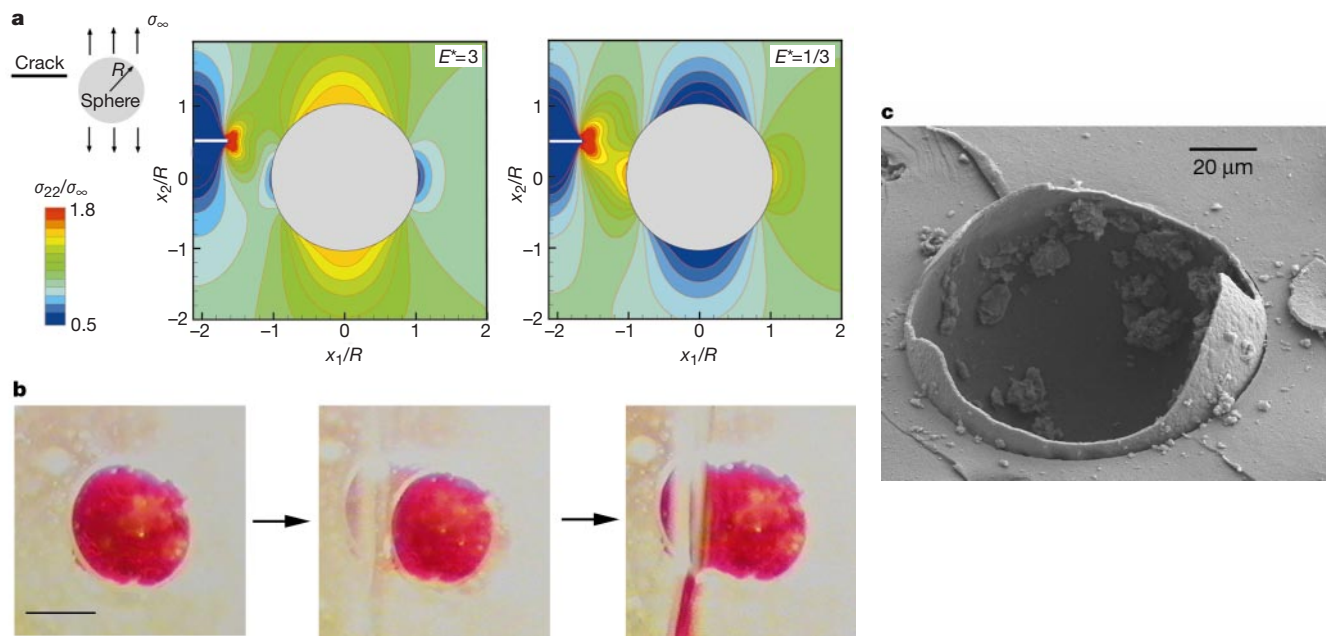


Figure 2 Rupture and release of the microencapsulated healing agent. **a**, Stress state in the vicinity of a planar crack as it approaches a spherical inclusion embedded in a linearly elastic matrix and subjected to a remote tensile loading perpendicular to the fracture plane. The left and right figures correspond to an inclusion three times stiffer ($E^* = E_{\text{sphere}}/E_{\text{matrix}} = 3$) and three times more compliant ($E^* = 1/3$) than the surrounding matrix, respectively. The Poisson's ratios of the sphere and matrix are equal

(0.30). **b**, A time sequence of video images shows the rupture of a microcapsule and the release of the healing agent. A red dye was added for visualization. The elapsed time from the left to right image is 1/15 s. Scale bar, 0.25 mm. **c**, A scanning electron microscope image shows the fracture plane of a self-healing material with a ruptured urea-formaldehyde microcapsule in a thermosetting matrix.

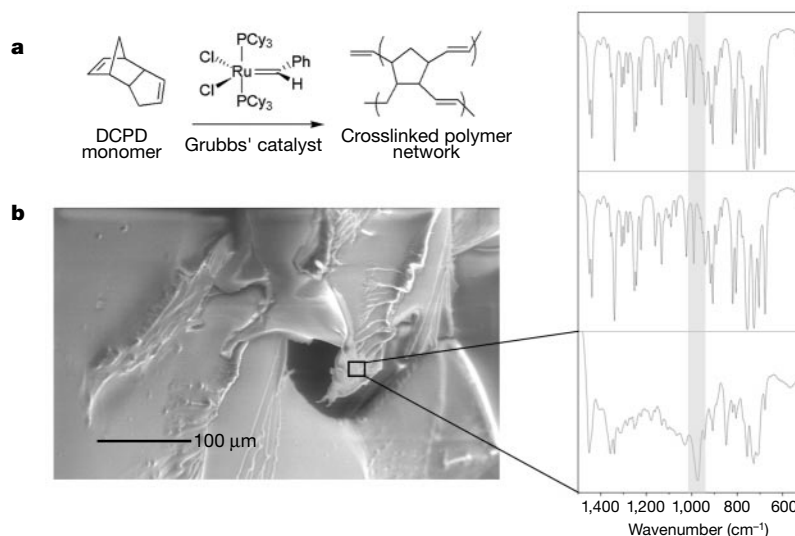


Figure 3 Chemistry of self-healing. **a**, Ruthenium-based Grubbs' catalyst initiates ring-opening metathesis polymerization (ROMP) of dicyclopentadiene (DCPD). **b**, Environmental ESEM micrograph and infrared analyses. IR spectra correspond to neat DCPD (top), an authentic sample of poly(DCPD) prepared with Grubbs' catalyst and DCPD

monomer (middle), and poly(DCPD) film formed at the healed interface (bottom). The highlighted peak at 965 cm^{-1} is characteristic of *trans* double bonds of ring-opened poly(DCPD).

microscope image of the fracture plane in Fig. 2c further illustrates the rupture process of an embedded microcapsule.

Completion of the self-healing process requires a suitable chemistry to polymerize the healing agent in the fracture plane. We identified the living ring-opening metathesis polymerization (ROMP) as meeting the diverse set of requirements of the self-healing system, which includes long shelf life, low monomer viscosity and volatility, rapid polymerization at ambient conditions, and low shrinkage upon polymerization. The ROMP reaction invokes the use of a transition metal catalyst (Grubbs' catalyst) that shows high metathesis activity while being tolerant of a wide range of functional groups as well as oxygen and water^{20–22}. The reaction polymerizes dicyclopentadiene (DCPD) at room temperature in several minutes to yield a tough and highly cross-linked polymer network (Fig. 3a). DCPD-filled microcapsules (50–200 μm) with a urea-formaldehyde shell were prepared using standard microencapsulation techniques. The microcapsule shell provides a protective barrier between the catalyst and DCPD to prevent polymerization during the preparation of the composite.

In order to test the stability and activity of the catalyst in an epoxy matrix, solution-state ^1H -NMR and solid-state ^{31}P -NMR spectro-

scopies of the self-healing polymer composite have been performed. Solutions of Grubbs' catalyst with the epoxy prepolymer and diethylenetriamine are chemically compatible as revealed by the characteristic benzylidene resonance at 23.3 p.p.m. Solid state ^{31}P -NMR measurements of a cured epoxy material with Grubbs' catalyst shows a characteristic signal corresponding to the tricyclohexylphosphine (PCy_3) coordinated to the ruthenium metal present in the catalyst. Solid state ^1H -NMR spectroscopy of the self-healing polymer composite also indicates the presence of a liquid DCPD monomer phase within the cured epoxy matrix.

Evidence of polymerization of the healing agent induced by damage is provided by environmental scanning electron microscopy (ESEM) and infrared spectroscopy (Fig. 3b). ESEM micrographs reveal the presence of a thin polymer film on the fracture surface. Infrared spectroscopy of this film indicates an absorption at 965 cm^{-1} characteristic of the ring-opened product, poly(DCPD). Control samples in which the catalyst was excluded showed no ability to polymerize the DCPD monomer, providing evidence that the embedded catalyst initiated the polymerization of the healing agent.

To assess the crack-healing efficiency of these composite materials, fracture tests were performed using a tapered double-cantilever beam (TDCB) specimen (Fig. 4). Self-healing composite and control samples were fabricated. Control samples consisted of: (1) neat epoxy containing no Grubbs' catalyst or microspheres; (2) epoxy with Grubbs' catalyst but no microspheres; and (3) epoxy with microspheres but no catalyst. A sharp pre-crack was created in the tapered samples by gently tapping a razor blade into a moulded starter notch. Load was applied in a direction perpendicular to the precrack (Mode I) with pin loading grips as shown in Fig. 4a. The virgin fracture toughness was determined from the critical load to propagate the crack and fail the specimen. After failure, the load was removed and the crack allowed to heal at room temperature with no manual intervention. Fracture tests were repeated after 48 hours to quantify the amount of healing and in all of the healed samples, the crack propagated along the original (virgin) crack plane. The intrinsic ability of the healing agent to rebond epoxy is shown by the upper horizontal dotted line in Fig. 4a which represents the average fracture load achieved for control samples (1) manually healed by injecting a mixture of DCPD and Grubbs' catalyst into the crack plane.

A representative load–displacement curve for a self-healing composite sample is plotted in Fig. 4a demonstrating recovery of about 75% of the virgin fracture load. In great contrast, all three types of control samples showed no healing and were unable to carry any load upon reloading. A set of four independently prepared self-healing composite samples showed an average healing efficiency of 60%. When the healing efficiency is calculated relative to the critical load for the virgin, neat resin control (lower horizontal dotted line in Fig. 4a), a value slightly greater than 100% is achieved. The average critical load for virgin self-healing samples containing microspheres and Grubbs' catalyst was 20% larger than the average value for the neat epoxy control samples, indicating that the addition of microspheres and catalyst increases the inherent toughness of the epoxy.

Self-healing composites possess great potential for solving some of the most limiting problems of polymeric structural materials: microcracking and hidden damage. Microcracks are the precursors to structural failure and the ability to heal them will enable structures with longer lifetimes and less maintenance. Filling microcracks will also mitigate the deleterious effects of environmentally assisted degradation such as moisture swelling and stress corrosion cracking. Although the potential benefits are quite high, the specific composite described here has some practical limitations on crack-healing kinetics and the stability of the catalyst to environmental conditions.

We have thus developed a new structural polymeric material that possesses the ability to heal cracks autonomically and recover

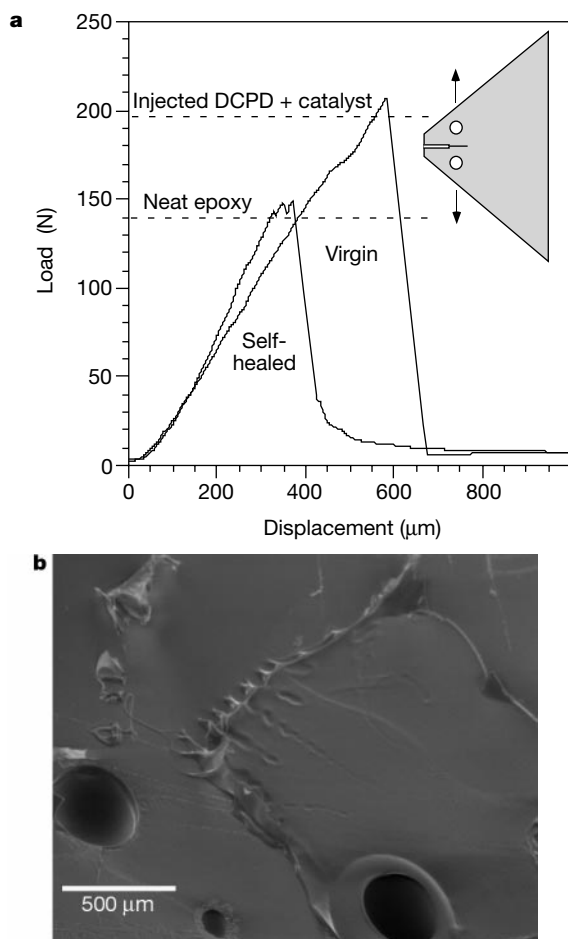


Figure 4 Self-healing efficiency in an epoxy polymer. **a**, Healing efficiency is obtained by fracture toughness testing of tapered double-cantilever beam (TDCB) specimens. The virgin fracture toughness is determined by propagating the starter crack along the mid-plane of the specimen. Subsequently, the load is removed and the crack allowed to heal at room temperature with no manual intervention. The healed fracture toughness is then measured by retesting the specimen. **b**, Post-fracture analysis of the specimens revealed that the healed crack failed in an interfacial manner from one side of the epoxy/poly(DCPD) interface to the other. The ESEM image in **b** shows one area of the fracture plane of a healed specimen in which the poly(DCPD) film is still attached to the interface on the right side of the image. The film that originally covered the interface on the left side of the image is found on the opposite mating surface of the specimen.

structural function. Such materials will increase the reliability and service life of thermosetting polymers used in a wide variety of applications ranging from microelectronics to aerospace. These concepts may also be applicable to a broad class of brittle materials including ceramics and glasses. We expect that the field of self-healing, although still in its infancy, will evolve beyond the method presented here until true biomimetic healing is achieved by incorporating a circulatory system that continuously transports the necessary chemicals and building blocks of healing to the site of damage. □

Methods

Preparation of microcapsules by *in situ* polymerization

In a 600 ml beaker we dissolved urea (0.11 mol, 7.0 g) followed by resorcinol (0.5 g) and ammonium chloride (0.5 g) in water (150 ml). A 5 wt.% solution of ethylene maleic anhydride copolymer (100 ml) was added to the reaction mixture and the pH of the reaction mixture was adjusted to 3.5 using 10% NaOH solution. The reaction mixture was agitated at 454 r.p.m. and to the stirred solution we added 60 ml of dicyclopentadiene to achieve an average droplet size of 200 μm . To the agitated emulsion was added 37% formaldehyde (0.23 ml, 18.91 g) solution and then the temperature of the reaction mixture was raised to 50 °C and maintained for 2 h. After 2 h, 200 ml of water was added to the reaction mixture. After 4 h, the reaction mixture was cooled to room temperature and the microcapsules were separated. The microcapsule slurry was diluted with an additional 200 ml of water and washed with water (3×500 ml). The capsules were isolated by vacuum filtration, and air dried. The yield was 80%. Their average size was 220 μm .

Self-healing epoxy specimen manufacture

The epoxy matrix composite was prepared by mixing 100 parts EPON 828 (Shell Chemicals Inc.) epoxide with 12 parts DETA (diethylenetriamine) curing agent (Shell Chemicals Inc.). Self-healing epoxy specimens were prepared by mixing 2.5% (by weight) Grubbs' catalyst and 10% (by weight) microcapsules with the resin mixture described above. The resin was then poured into silicone rubber moulds and cured for 24 h at room temperature, followed by postcuring at 40 °C for 24 h.

TDCB specimens

The TDCB sample was introduced by Mostovoy and co-workers²³ and is designed so that the compliance of the specimen changes linearly with crack length during the test. This tapered geometry enables controlled crack growth across the centre of a brittle sample such as epoxy. The fracture toughness for TDCB specimens depends only on the applied load and is independent of the crack length so that $K_{IC} = \alpha P_c$ where α is a function of geometry and material properties and P_c is the critical load at fracture. K_{IC} is the experimentally determined mode-I critical stress intensity factor. A taper angle of 40° was used and α was measured to be $7,700 \text{ m}^{-3/2}$.

Quantifying healing efficiency

Previous studies of crack healing in thermoplastic polymers quantified healing effects by comparing the fracture toughness of the virgin material to the fracture toughness measured after crack closure and healing.^{13–16} An efficiency of healing was defined as the ratio of the fracture toughness of healed and virgin materials such that $\eta = K_{IC}^{\text{healed}}/K_{IC}^{\text{virgin}}$ where η is the healing efficiency.

Received 15 August; accepted 12 December 2000.

1. Talrega, R. Damage development in composites: mechanisms and modelling. *J. Strain Anal. Eng. Des.* **24**, 215–222 (1989).
2. Talrega, R. (ed.) *Damage Mechanics of Composite Materials* 139–241 (Elsevier, New York, 1994).
3. Gamstedt, E. K. & Talrega, R. Fatigue damage mechanisms in unidirectional carbon-fibre-reinforced plastics. *J. Mater. Sci.* **34**, 2535–2546 (1999).
4. Pecht, M. G., Nguyen, L. T. & Hackim, E. B. *Plastic-Encapsulated Microelectronics* 235–301 (John Wiley & Sons, New York, 1995).
5. Lee, L. H. *Adhesive Bonding* 239–291 (Plenum, New York, 1991).
6. Wool, R. P. *Polymer Interfaces: Structure and Strength* Ch. 12 445–479 (Hanser Gardner, Cincinnati, 1995).
7. Dry, C. & Sottos, N. in *Smart Structures and Materials 1993: Smart Materials* (ed. Varadan, V. K.) Vol. 1916, 438 (SPIE Proceedings, SPIE, Bellingham, WA, 1993).
8. Dry, C. Procedures developed for self-repair of polymeric matrix composite materials. *Comp. Struct.* **35**, 263–269 (1996).
9. Wiederhorn, S. M. & Townsend, P. R. Crack healing in glass. *J. Am. Ceram. Soc.* **53**, 486–489 (1970).
10. Stavrinidis, B. & Holloway, D. G. Crack healing in glass. *Phys. Chem. Glasses* **24**, 19–25 (1983).
11. Inagaki, M., Urashima, K., Toyomasu, S., Goto, Y. & Sakai, M. Work of fracture and crack healing in glass. *J. Am. Ceram. Soc.* **68**, 704–706 (1985).
12. Jud, K. & Kausch, H. H. Load transfer through chain molecules after interpenetration at interfaces. *Polym. Bull.* **1**, 697–707 (1979).
13. Jud, K., Kausch, H. H. & Williams, J. G. Fracture mechanics studies of crack healing and welding of polymers. *J. Mater. Sci.* **16**, 204–210 (1981).
14. Kausch, H. H. & Jud, K. Molecular aspects of crack formation and healing in glassy polymers. *Plastic Rubber Proc. Appl.* **2**, 265–268 (1982).
15. Wool, R. P. & O'Connor, K. M. A theory of crack healing in polymers. *J. Appl. Phys.* **52**, 5953–5963 (1982).
16. Wang, E. P., Lee, S. & Harmon, J. Ethanol-induced crack healing in poly(methyl methacrylate). *J. Polym. Sci. B* **32**, 1217–1227 (1994).

17. Lin, C. B., Lee, S. & Liu, K. S. Methanol-induced crack healing in poly(methyl methacrylate). *Polym. Eng. Sci.* **30**, 1399–1406 (1990).
18. Raghavan, J. & Wool, R. P. Interfaces in repair, recycling, joining and manufacturing of polymers and polymer composites. *J. Appl. Polym. Sci.* **71**, 775–785 (1999).
19. Mura, T. *Micromechanics of Defects in Solids* 2nd edn (Kluwer Academic, New York, 1987).
20. Grubbs, R. H. & Tumas, W. Polymer synthesis and organotransition metal chemistry. *Science* **243**, 907–915 (1989).
21. Schwab, P., Grubbs, R. H. & Ziller, J. W. Synthesis and applications of $\text{RuCl}_2(\text{=CHR})(\text{PR}_3)_2$: the influence of the alkylidene moiety on metathesis activity. *J. Am. Chem. Soc.* **118**, 100–110 (1996).
22. Sanford, M. S., Henling, L. M. & Grubbs, R. H. Synthesis and reactivity of neutral and cationic ruthenium(II) tri(pyrazolyl)borate alkylidenes. *Organometallics* **17**, 5384–5389 (1998).
23. Mostovoy, S., Croseley, P. B. & Ripling, E. J. Use of crack-line-loaded specimens for measuring plane-strain fracture toughness. *J. Mater.* **2**, 661–681 (1967).

Acknowledgements

This work has been sponsored by the University of Illinois Critical Research Initiative Program and the AFOSR Aerospace and Materials Science Directorate. Electron microscopy was carried out in the Center for Microanalysis of Materials, University of Illinois, which is supported by the US Department of Energy.

Correspondence and requests for materials should be addressed to S.W. (e-mail: swhite@uiuc.edu).

A chiroselective peptide replicator

Alan Saghatelian, Yohei Yokobayashi, Kathy Soltani & M. Reza Ghadiri

Departments of Chemistry and Molecular Biology and the Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, California 92037, USA

The origin of homochirality in living systems is often attributed to the generation of enantiomeric differences in a pool of chiral prebiotic molecules^{1,2}, but none of the possible physicochemical processes considered^{1–7} can produce the significant imbalance required if homochiral biopolymers are to result from simple coupling of suitable precursor molecules. This implies a central role either for additional processes that can selectively amplify an initially minute enantiomeric difference in the starting material^{11,8–12}, or for a nonenzymatic process by which biopolymers undergo chiroselective molecular replication^{13–16}. Given that molecular self-replication and the capacity for selection are necessary conditions for the emergence of life, chiroselective replication of biopolymers seems a particularly attractive process for explaining homochirality in nature^{13–16}. Here we report that a 32-residue peptide replicator, designed according to our earlier principles^{17–20}, is capable of efficiently amplifying homochiral products from a racemic mixture of peptide fragments through a chiroselective autocatalytic cycle. The chiroselective amplification process discriminates between structures possessing even single stereochemical mutations within otherwise homochiral sequences. Moreover, the system exhibits a dynamic stereochemical 'editing' function; in contrast to the previously observed error correction²⁰, it makes use of heterochiral sequences that arise through uncatalysed background reactions to catalyse the production of the homochiral product. These results support the idea that self-replicating polypeptides could have played a key role in the origin of homochirality on Earth.

Chiroselective amplification refers to an autocatalytic process in which a homochiral template instructs the synthesis of a homochiral polymer of the same handedness¹³. Past efforts aimed at establishing the feasibility of nonenzymatic chiroselective amplification, using nucleic acids and their analogues, have been hampered by lack of template turnover¹³ and/or enantiomeric cross-inhibition processes in template-directed oligomerization of activated monomeric building blocks^{14–16}.

The peptide used in the present study was identified using design and mechanistic principles that are similar to previously

described systems^{17–20}. An enantiomeric pair of electrophilic E and nucleophilic N peptide fragments was employed in order to probe the relationship between self-replication and homochirality (Fig. 1). Each pair of electrophile or nucleophile contains one member that is composed of entirely L amino acids, N^L and E^L, and the other contains only unnatural D amino acids, N^D and E^D. Reactions between these four substrates can give rise to four possible products: homochiral templates, T^{LL} and T^{DD}, or heterochiral products, T^{LD} and T^{DL} (Fig. 2).

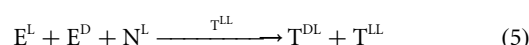
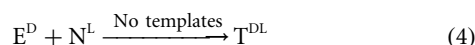
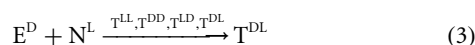
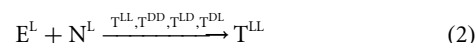
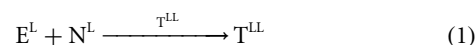
From a 'racemic' starting mixture made up of equal amounts of electrophilic (E^D and E^L) and nucleophilic (N^L and N^D) fragments, homochiral products (T^{LL} and T^{DD}) are preferentially produced (Fig. 3a). Product formation and diminution of starting materials were monitored by reverse-phase high-performance liquid chromatography (RP-HPLC) and validated by mass spectrometry and HPLC retention time comparisons with authentic materials. The observed generation of increasing diastereomeric excess during the course of the reaction is due to the autocatalytic activity of the homochiral templates, which initially appear in the reaction mixture along with the heterochiral products through uncatalysed background coupling reactions (Fig. 3b). To study the mechanism of chiroselective amplification, we systematically analysed the various steps involved in product formation.

First we examined the degree of inherent stereoselectivity in the fragment condensation reactions²¹ which can be studied only in the absence of template control (interhelical noncovalent interactions) and other biasing structural factors. Thus a 'racemic' mixture of the four fragments were allowed to react in buffered solutions containing 3 M GndHCl to ensure denaturation of catalytically

competent intermolecular complexes.

Both homochiral and heterochiral products were produced at the same rate, indicating that the ligation chemistry was not inherently diastereoselective (Fig. 3a). Therefore, the selective amplification of homochiral products under native (non-denaturing) reaction conditions is the direct result of chiroselective autocatalysis (Fig. 2).

To further investigate the chiroselectivity and examine the possibility of cross-isomer inhibition, two sets of reactions were performed where the only possible product was either homochiral or heterochiral (reactions 1–4). In one series, we examined the possible catalytic, autocatalytic, and/or inhibitory effects of the four products on the rate of T^{LL} production (reaction 2).



T^{LL} autocatalytically accelerates its own production in reaction mixtures containing equimolar amounts of N^L and E^L (reaction 1; Fig. 3c), and adding T^{DD}, T^{DL} or T^{LD} individually did not have any observable influence on the rate of T^{LL} production. Indeed, reactions between N^L and E^L in the presence of equimolar amounts of all four templates, T^{LL}, T^{DD}, T^{DL} and T^{LD} (reaction 2), displayed a

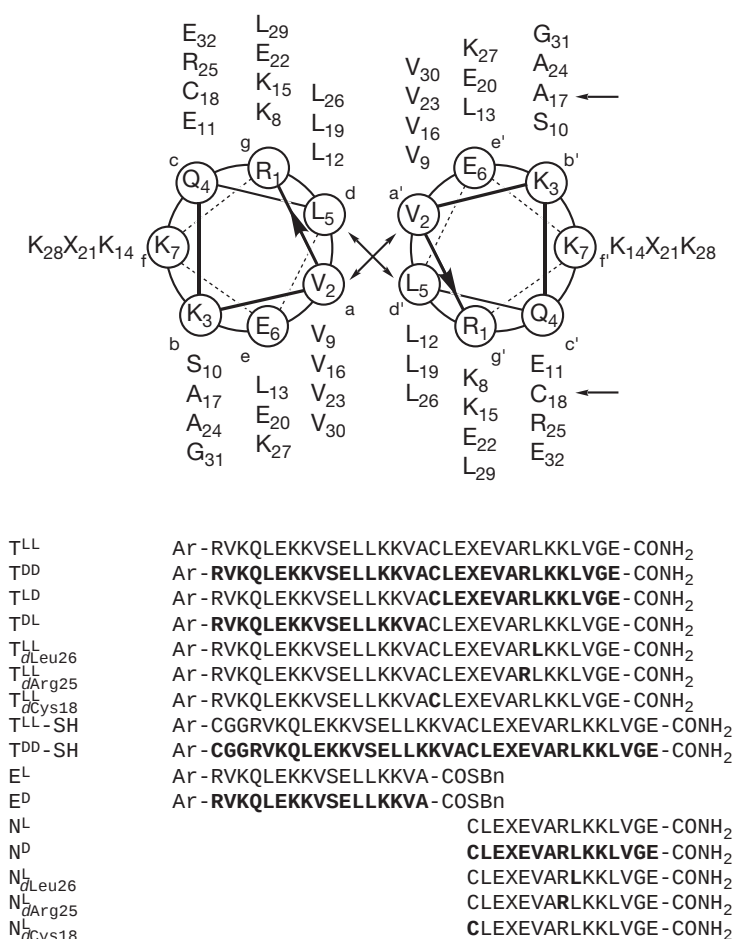


Figure 1 Helical wheel diagram and peptide sequences employed in this study. The chiroselective peptide replicator is shown in a homodimeric coiled-coil state with arrows representing the ligation site at the solvent-exposed helical surface. Amino acid residues

at the a, d, e and g positions compose the interhelical recognition interface. Amino acid residues with D configuration are shown in bold. Bn, benzyl; Ar, 4-acetamidobenzoyl; and X, lysine-ε-NHCO-Ar.

similar rate of product formation to that of the reaction where T^{LL} was the only template present (Fig. 3c). These experiments suggest that T^{LL} is the only active template involved in the ligation of E^L and N^L , and that all the other templates (enantiomer and diastereomers) act only as spectators during the formation of T^{LL} (Fig. 3c).

When the fragments N^L and E^D are reacted to study the production of the heterochiral template T^{LD} , the reaction rates are largely the same as in the presence (reaction 3) and absence (reaction 4) of

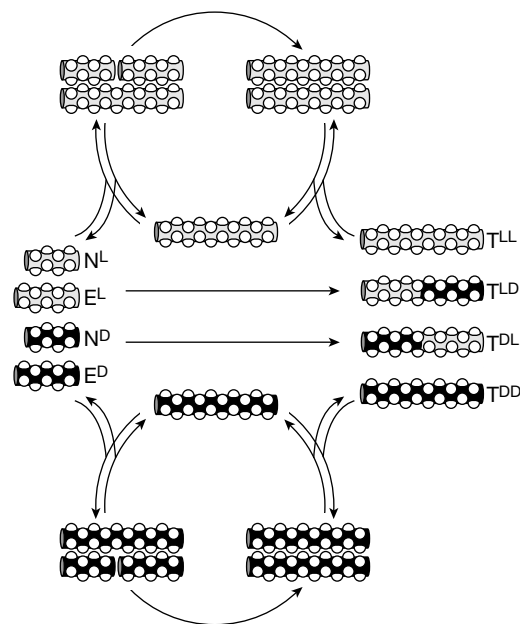


Figure 2 A schematic representation of the chiroselective replication cycles. Homochiral peptides T^{LL} and T^{DD} are produced autocatalytically^{17,18} while the heterochiral peptides T^{LD} and T^{DL} result from uncatalysed background fragment condensation reactions²⁷. Template-directed ligation reactions proceed through the intermediary of stereospecific noncovalent complexes (ternary and/or quaternary) and pass the stereochemical information from the homochiral products to the substrates, thus resulting in the amplification of homochiral products (for clarity only ternary complexes are depicted). Peptide fragments are shown in helical conformation and represented as cylinders. Light and dark backgrounds denote regions of the sequence composed of L- and D-amino acids, respectively.

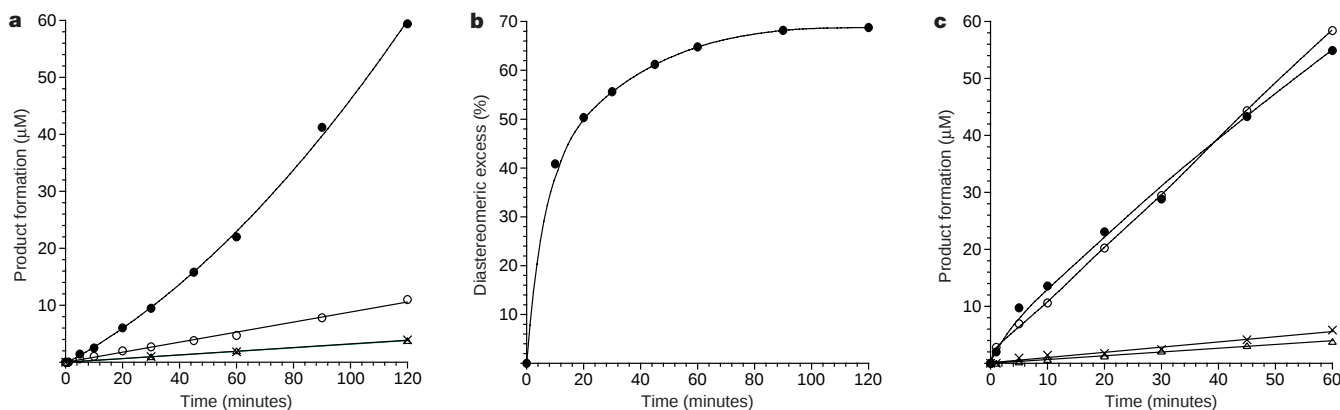


Figure 3 Rates of product formation with time. **a**, Template-directed chiroselective amplification. Production of homochiral products T^{LL} and T^{DD} (filled circles) and heterochiral products T^{LD} and T^{DL} (empty circles) is shown as a function of time for a reaction mixture containing 100-μM concentrations of each of fragments N^L , N^D , E^D and E^L . Production rates are shown for homochiral (empty triangle) and heterochiral products (cross) in a similar reaction mixture but in the presence of 3 M GndHCl. **b**, Increasing diastereomeric excess $100 \times [(T^{LL} + T^{DD}) - (T^{LD} + T^{DL})]/T^{total}$ as a function of time signifying amplification of homochiral products (data taken from Fig. 3a). **c**, Absence of enantiomeric cross-inhibition in chiroselective peptide replication and (auto)catalytic

any or all of the four templates (Fig. 3c). These experiments indicate that the heterochiral products are produced in a template-independent fashion largely through uncatalysed bimolecular background reactions. Thus the homochiral templates are highly chiroselective autocatalysts that neither inhibit nor promote the ligation of the other products studied.

The possibility of substrate-mediated enantiomeric cross-inhibition^{14–16} was examined by comparing the rates of T^{LL} formation in reactions 1 and 5 under similar reaction conditions. Analyses of reaction mixtures containing equimolar amounts of E^L , N^L and 30 mole% T^{LL} either in the absence (reaction 1) or presence of equimolar amounts of E^D (reaction 5) gave similar rates of T^{LL} production, thus demonstrating the absence of enantiomeric cross-inhibition in the self-replication process.

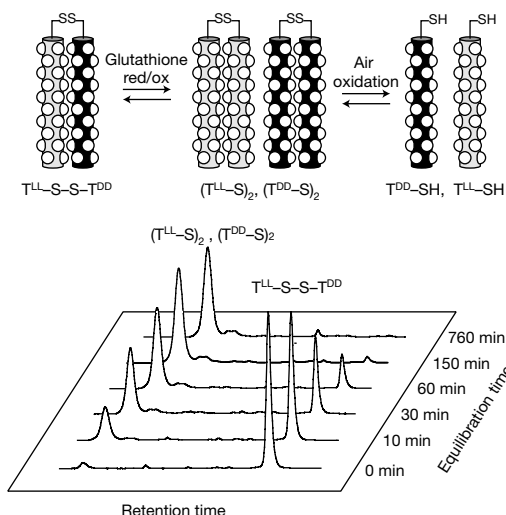


Figure 4 Schematic representation of chiroselective coiled-coil formation processes and the corresponding experimental HPLC time traces. Preformed heterochiral disulphide dimer T^{DD} -S- T^{LL} (21.5 μM) in a glutathione redox buffer (500 μM reduced glutathione, 125 μM oxidized glutathione, 2 M NaCl, 100 mM MOPS buffer pH 7.5) undergoes complete disproportionation into homochiral assemblies T^{LL} -S- T^{LL} and T^{DD} -S- T^{DD} , as indicated by the RP-HPLC time traces, which suggests that coiled-coil formation is intrinsically chiroselective.

inactivity of heterochiral templates. Production of homochiral product T^{LL} in a reaction mixture containing 100 μM concentration of each fragment E^L and N^L is shown, in the presence of 25 μM initial concentration of T^{LL} (open circles) or in the presence of 25 μM initial concentrations of each of the products T^{LL} , T^{DD} , T^{LD} and T^{DL} (filled circles). Production rates of heterochiral product T^{DL} in a reaction mixture containing 100 μM concentration of fragments E^D and N^L are shown, in the presence (open triangle) or absence (cross) of 25 μM initial concentration of T^{LL} , T^{DD} , T^{LD} and T^{DL} . The lines are solely intended as a visual guide. All reactions, unless noted, were performed in 100 mM MOPS, 2 M NaCl, 1% v/v BnSH, pH 7.5 as described previously^{17,18}. MOPS, 3-(N-morpholino)-propanesulphonic acid.

To investigate the thermodynamic basis of the chiroselective peptide self-replication process, the stereoselectivity in coiled-coil dimer formation²² was studied by employing two enantiomeric homochiral templates, T^{DD} -SH and T^{LL} -SH. These peptides differ in sequence from T^{LL} and T^{DD} by only a Cys18Ser mutation and a Gly-Gly-Cys appendage at the amino terminus. The flexible appendage was employed to allow the formation of disulphide linked coiled coils²³ and the Cys18Ser mutation to ensure regioselectivity in disulphide bond-forming reactions (see below). The stereospecificity of coiled-coil formation was studied by allowing the racemic mixture of T^{DD} -SH and T^{LL} -SH to undergo slow air oxidation (25 μ M each fragment in 100 mM MOPS, 2 M NaCl, pH 7.5) or by promoting disulphide exchange with preformed heterochiral disulphide-bonded dimer T^{DDS} - ST^{LL} in a glutathione redox buffer (Fig. 4a). HPLC analyses show that both the racemic mixture of T^{DD} -SH and T^{LL} -SH and the heterochiral dimer, T^{DDS} - ST^{LL} , are cleanly and quantitatively converted to the homochiral dimers, T^{LLS} - ST^{LL} and T^{DDS} - ST^{DD} (Fig. 4b), showing the same selectivity seen in the autocatalytic reactions. These results support the notion that the selectivity in autocatalytic fragment condensation is driven by the same homochiral molecular recognition events that favour the formation of the homochiral coiled-coils.

The limits of chiroselective amplification were tested by probing

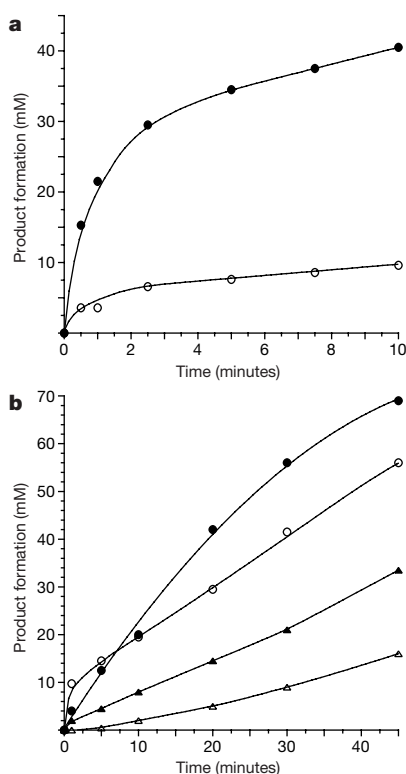


Figure 5 Rates of production formation in time. **a**, Probing the limits of chiroselective amplification through selective stereochemical mutations. Production of homochiral product T^{LL} (filled circles) and heterochiral product T^{LL}_{dCys18} (open circle) are shown as a function of time in a reaction mixture containing equimolar (100 μ M) concentration of competing fragments N^L and N^L_{dCys18} and 50 μ M concentrations of each of E^L , T^{LL} , and T^{LL}_{dCys18} (production rates were quantitatively monitored by HPLC analyses of the depletion of N^L and N^L_{dCys18} ; owing to product peak overlaps). **b**, Heterochiral mutant products display a dynamic stereochemical editing function by cross-catalysing the production of the homochiral replicator sequence in a template-dependent fashion. Production of homochiral product T^{LL} in reaction mixtures containing 100 μ M concentrations of N^L and E^L is efficiently catalysed in the presence of 50 μ M initial concentrations of T^{LL}_{dLeu26} (filled circle), T^{LL}_{dCys18} (open circle) or T^{LL}_{dArg25} (filled triangle). For comparison, the rate of production of T^{LL} in a similar reaction mixture but in the absence of added initial template is shown (open triangle). All reactions, unless noted, were performed in 100 mM MOPS, 2 M NaCl, 2 mM tris(carboxyethyl) phosphine, pH 7.5, as described previously^{17,18}.

the effects of single amino acid stereochemical mutations within otherwise homochiral peptide fragments. Three nucleophilic peptide fragments N^L_{dLeu26} , N^L_{dArg25} and N^L_{dCys18} were prepared, each bearing a single D-amino acid substitution within the 15-residue homochiral nucleophilic fragment N^L (Fig. 1). The mutants N^L_{dLeu26} and N^L_{dArg25} were designed to test the effects of a stereochemical mutation within the informational complementary hydrophobic recognition interface and the non-informational solvent-exposed helical surface, respectively. Reaction mixtures containing equimolar amounts of E^L and either of the mutants N^L_{dLeu26} or N^L_{dArg25} produced only background rates of the corresponding fragment condensation products T^{LL}_{dLeu26} and T^{LL}_{dArg25} , either in the presence or absence of 25 mole% of the template T^{LL} . Moreover, neither T^{LL}_{dLeu26} nor T^{LL}_{dArg25} were autocatalytically active. The third peptide fragment N^L_{dCys18} was designed to provide the most stringent test of chiroselectivity, because changing the chirality of the N-terminal residue was expected to have a minimal effect on the stability of the nascent helical fragment in solution or on its interactions with the template. Reaction mixtures containing equimolar amounts of E^L and N^L_{dCys18} and 50 mole% of either T^{LL} or T^{LL}_{dCys18} yielded a small rate enhancement in the formation of product T^{LL}_{dCys18} (1.3 times over the background). However, in a direct competition experiment employing a 100 μ M of N^L and N^L_{dCys18} , 50 μ M of E^L , and 50 μ M each of T^{LL} and T^{LL}_{dCys18} , the homochiral peptide T^{LL} quickly became the dominant species in the reaction mixture owing to its faster rate of production (Fig. 5a). The above studies establish that chiroselectivity in peptide self-replication is remarkably robust, not favouring reactants and products that differ stereochemically by one amino acid out of 15 or 32 residues.

An unexpected result of the above studies was that the three single stereochemical mutant products T^{LL}_{dLeu26} , T^{LL}_{dArg25} and T^{LL}_{dCys18} are efficient catalysts for the template-directed production of T^{LL} (Fig. 5b). A similar effect was observed previously²⁰ and shown to be the result of an autocatalytic self-organized network through which closely related mutant peptides ('errors' produced through background fragment condensation reactions) act as specific catalysts for the production of the native replicator (see ref. 20 for a discussion of the thermodynamic factors that may cause such an effect). In the present case, the heterochiral mutants T^{LL}_{dLeu26} , T^{LL}_{dArg25} and T^{LL}_{dCys18} accelerate the rate of production of the homochiral replicator T^{LL} by coupling the mutant template-mediate cross-catalytic cycles to the

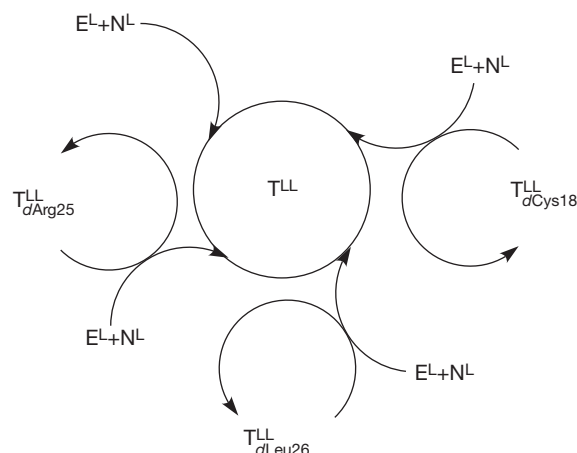


Figure 6 Schematic illustration of cross-catalytic networks established by heterochiral templates that help in the maintenance and amplification of homochiral products. The scheme depicts at the centre the autocatalytic cycle of homochiral replicator T^{LL} fed by the catalytic actions of three heterochiral single stereochemical mutant templates T^{LL}_{dLeu26} , T^{LL}_{dCys18} , and T^{LL}_{dArg25} . Overall, the system displays a dynamic stereochemical editing function as a result of the appearance of heterochiral sequences in the reaction mixture. For clarity, formation of mutant templates through uncatalysed background fragment condensation reactions is not shown.

autocatalytic production cycle of the homochiral template (Fig. 6). Cross-catalysis seems to result from the ability of these peptides to tolerate D-amino acid substitutions while maintaining a catalytically active conformation^{24–26}. We refer to this process as ‘dynamic stereochemical editing’.

Thus we have shown that even within the simple helical structure, a peptide replicator can contain a rich degree of information that enables it to display, at the molecular level, some of the basic characteristics of living systems, such as self-replication, homochirality, and resistance toward accumulation of errors (stereochemical mutations). The link between these traits allows the replicator to selectively amplify homochiral sequences. Therefore, chiroselectivity in peptide self-replication is a direct result of complementary noncovalent interactions that pass on both binding and stereochemical information simultaneously. That such a prototypical peptide system is able to amplify homochiral products through self-replication suggests that this and similar mechanisms may have affected the origin of homochirality on Earth. □

Received 19 October; accepted 8 December 2000.

- Bonner, W. A. The origin and amplification of biomolecular chirality. *Origins Life Evol. Biosph.* **21**, 59–111 (1991).
- Bada, J. L. Origins of homochirality. *Nature* **374**, 594–595 (1995).
- Bonner, W. A., Blair, N. E. & Dribas, F. M. Experiments on the abiotic amplification of optical activity. *Origins Life* **11**, 119–134 (1981).
- Cronin, J. R. & Pizzarello, S. Enantiomeric excesses in meteoritic amino acids. *Science* **275**, 951–955 (1997).
- Kondepudi, D. K. & Nelson, G. W. Weak neutral currents and the origin of biomolecular chirality. *Nature* **314**, 438–441 (1985).
- Kondepudi, D. K., Kaufman, R. J. & Singh, N. Chiral symmetry breaking in sodium chlorate crystallization. *Science* **250**, 975–977 (1990).
- Weissbuch, I. et al. Spontaneous generation and amplification of optical activity in α -amino acids by enantioselective occlusion into centrosymmetric crystals of glycine. *Nature* **310**, 161–164 (1984).
- Frank, F. C. On spontaneous asymmetric synthesis. *Biochim. Biophys. Acta* **11**, 459–463 (1953).
- Calvin, M. *Chemical Evolution* (Oxford Univ. Press, Oxford, 1969).
- Girdard, C. & Kagan, H. B. Nonlinear effects in asymmetric synthesis and stereoselective reactions: ten years of investigation. *Angew. Chem. Int. Edn Engl.* **37**, 2923–2959 (1998).
- Soai, K., Shibata, T., Morioka, H. & Choji, K. Asymmetric autocatalysis and amplification of enantiomeric excess of a chiral molecule. *Nature* **378**, 767–768 (1995).
- Siegel, J. S. Homochiral imperative of molecular evolution. *Chirality* **10**, 24–27 (1998).
- Bolli, M., Micura, R. & Eschenmoser, A. Pyranosyl-RNA: chiroselective self-assembly of base sequences by ligative oligomerization of tetranucleotide-2',3'-cyclophosphates (with a commentary concerning the origin of biomolecular homochirality). *Chem. Biol.* **4**, 309–320 (1997).
- Joyce, G. F. et al. Chiral selection in poly(C)-directed synthesis of oligo(G). *Nature* **310**, 602–604 (1984).
- Schmidt, J. G., Nielsen, P. E. & Orgel, L. E. Enantiomeric cross-inhibition in the synthesis of oligonucleotides on a nonchiral template. *J. Am. Chem. Soc.* **119**, 1494–1495 (1997).
- Kozlov, I. A., Politis, P. K., Pitsch, S., Herdewijn, P. & Orgel, L. E. A highly enantio-selective hexitol nucleic acid template for nonenzymic oligonucleotide synthesis. *J. Am. Chem. Soc.* **121**, 1108–1109 (1999).
- Lee, D. H., Granja, J. R., Martinez, J. A., Severin, K. & Ghadiri, M. R. A self-replicating peptide. *Nature* **382**, 525–528 (1996).
- Severin, K., Lee, D. H., Martinez, J. A. & Ghadiri, M. R. Peptide self-replication via template-directed ligation. *Chem. Eur. J.* **3**, 1017–1024 (1997).
- Lee, D. H., Severin, K., Yokobayashi, Y. & Ghadiri, M. R. Emergence of symbiosis in peptide self-replication through a hypercyclic network. *Nature* **390**, 591–594 (1997).
- Severin, K., Lee, D. H., Martinez, J. A., Vieth, M. & Ghadiri, M. R. Dynamic error correction in autocatalytic peptide networks. *Angew. Chem. Int. Edn Engl.* **37**, 126–128 (1998).
- Li, T., Budt, K. H. & Kishi, Y. Influence of secondary structure (helical conformation) on stereoselectivity in peptide couplings. *J. Chem. Soc. Chem. Commun.* **24**, 1817–1819 (1987).
- Case, M. A., Ghadiri, M. R., Mutz, M. W. & McLendon, G. L. Stereoselection in designed three-helix bundle metalloproteins. *Chirality* **10**, 35–40 (1998).
- Harbury, P. B., Zhang, T., Kim, P. S. & Alber, T. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401–1407 (1993).
- Krause, E., Bienert, M., Schmieder, P. & Wenshub, H. The helix-destabilizing propensity scale of D-amino acids: the influence of side chain steric effects. *J. Am. Chem. Soc.* **122**, 4865–4870 (2000).
- Rothmund, S. et al. Structure effects of double D-amino acid replacements: A nuclear magnetic resonance and circular dichroism study using amphipathic model helices. *Biochemistry* **34**, 12954–12962 (1995).
- Rothmund, S. et al. Peptide destabilization by two adjacent D-amino acids in single-stranded amphipathic α -helices. *Pept. Res.* **9**, 78–87 (1996).
- Dawson, P. E., Muir, T. W., Clark-Lewis, I. & Kent, S. B. H. Synthesis of proteins by native chemical ligation. *Science* **266**, 776–779 (1994).

Acknowledgements

We thank our colleagues A. Eschenmoser and L. Orgel for their helpful suggestions and critical review of this manuscript, and the NASA Astrobiology Institute for financial support.

Correspondence and requests for materials should be addressed to M.R.G. (e-mail: ghadiri@scripps.edu).

Evidence for non-selective preservation of organic matter in sinking marine particles

John I. Hedges*, Jeffrey A. Baldock†, Yves G  linas*, Cindy Lee‡, Michael Peterson* & Stuart G. Wakeham  

* School of Oceanography, Box 357940, University of Washington, Seattle, Washington 98195-7940, USA

† CSIRO Land and Water, PMB #2, Glen Osmond, South Australia 5064, Australia

‡ Marine Sciences Research Center, State University of New York, Stony Brook, New York 11794, USA

   Skidaway Institute of Oceanography, 10 Ocean Science Circle, Savannah, Georgia 31411, USA

The sinking of particulate organic matter from ocean surface waters transports carbon to the ocean interior^{1,2}, where almost all is then recycled. The unrecycled fraction of this organic matter can become buried in ocean sediments, thus sequestering carbon and so influencing atmospheric carbon dioxide concentrations³. The processes controlling the extensive biodegradation of sinking particles remain unclear, partly because of the difficulty in resolving the composition of the residual organic matter at depth with existing chromatographic techniques⁴. Here, using solid-state ¹³C NMR spectroscopy⁵, we characterize the chemical structure of organic carbon in both surface plankton and sinking particulate matter from the Pacific Ocean⁴ and the Arabian Sea⁶. We found that minimal changes occur in bulk organic composition, despite extensive (>98%) biodegradation, and that amino-acid-like material predominates throughout the water column in both regions. The compositional similarity between phytoplankton biomass and the small remnant of organic matter reaching the ocean interior indicates that the formation of unusual biochemicals, either by chemical recombination⁷ or microbial biosynthesis⁸, is not the main process controlling the preservation of particulate organic carbon within the water column at these two sites. We suggest instead that organic matter might be protected from degradation by the inorganic matrix of sinking particles.

The ocean interior acts as a global heterotrophic digester, where almost all organic matter injected from surface (0–100 m) waters is respired to inorganic nutrients⁹. Such ‘remineralization’ is important because sinking organic particles selectively transport bioactive elements through density-stratified water columns typical of temperate oceans¹. Although the size and density of sinking particles affect nutrient release at depth, reaction rates of the component biochemicals are also important considerations^{9,10}. Unfortunately, the percentage of the total organic matter in this particle rain that can be molecularly characterized by conventional chromatographic analysis decreases rapidly down ocean water columns to become a minor component (<25%) at depth⁴. The attenuation of definitive molecular information with increasing ocean depth¹¹ greatly diminishes our understanding of the forms and reaction potentials of organic materials flowing through one of the most ecologically important, and climatically critical, transport pathways in the ocean¹².

Several hypotheses have been offered for the universally observed decrease in molecularly resolvable organic substances attending advanced biodegradation. The classical explanation is that labile intermediates released during microbial degradation of biomacromolecules (for example, polysaccharides, lignins and proteins) spontaneously recombine to form chemically complex ‘geopolymers’⁷. Products of such abiotic ‘humification’ reactions

may be too structurally complex to chemically analyse or enzymatically degrade. Alternatively, organisms may synthesize biomacromolecules that are intrinsically resistant to biodegradation and which become 'selectively preserved' in soils and sediments where they resist conventional analysis⁸. Numerous hydrolysis-resistant biomacromolecules have now been identified in terrestrial and marine plants and bacteria, as well as in modern and ancient sediments¹³. Finally, it has been postulated that refractory organic or inorganic matrices may shield intrinsically labile organic substances so that these 'armoured' substrates might persist much longer than their chemistries would suggest¹⁴. An important contrast among these protection mechanisms is that humification and selective preservation should be attended by pronounced changes in organic structure, whereas shielding does not require large compositional changes.

We tested these hypotheses by using solid-state ¹³C nuclear magnetic resonance (NMR) spectroscopy to characterize the chemistry of organic carbon contained in bulk samples of plankton (0–100 m depth) and particles sinking through upper (~500–1,000 m) and lower (~3,000–3,500 m) water columns of the equatorial Pacific Ocean and the western Arabian Sea (Table 1). The sinking particle samples from the Pacific Ocean were collected over a year (January 1991–January 1992) in biocide-treated sediment traps¹⁵ at an equatorial ocean site characterized by moderate productivity, low seasonality and moderate (100–200 µmol l⁻¹) dissolved O₂ concentrations in interior waters¹⁶. Sinking particles were similarly collected in the Arabian Sea over a seven-month period (May–December 1995) at a site that experiences high productivity during the southwest monsoon and exhibits low (<10 µmol l⁻¹) dissolved O₂ concentrations below the upper ~200 m (ref. 17). Vertical fluxes of particulate organic carbon (POC) down both water columns decrease by roughly two orders of magnitude from POC produced by net primary production^{6,15,18}, exemplifying extensive degradation of particles injected by the biological pump in these contrasting ocean regimes (Table 1). Previous chromatographic analyses of more than 100 amino acids, carbohydrates and lipids in samples collected from the Pacific site demonstrated that molecularly uncharacterized POC increased progressively from approximately 20% in plankton to 75% in the deep water column⁴. Although the present study is based on samples collected over a year or less from only two ocean sites, their flux attenuations and organic compositions are typical of many other open ocean regions⁴.

To the best of our knowledge, this is the first use of solid-state ¹³C NMR to characterize the chemical structure of organic carbon in particles sinking through the ocean. This technique is well suited for such characterizations, because most organic-rich samples such as plankton and sediment-trap material can be analysed with no preparation other than drying and grinding⁵. With careful attention to spectral acquisition techniques¹⁹, a representative cross-section of all forms of organic carbon in the solid sample can be detected using solid-state ¹³C NMR, regardless of whether it exists in a free or shielded form. Because solid-state ¹³C NMR 'sees' both the molecularly measured and the molecularly uncharacterized fractions of bulk particulate materials, it directly addresses the central question

of whether these two fractions are compositionally similar. This issue is most critical in deeper water columns, where the extent of organic matter degradation is most pronounced and the fraction of chromatographically resolvable molecules is small.

The solid-state ¹³C-NMR spectra of the plankton and sinking particle mixtures from the Pacific Ocean and Arabian Sea sites are remarkably similar with depth and between sites, considering that the deep sediment-trap samples represent less than 1% of the local net primary production (Fig. 1 and Table 1). The largest compositional change is the 9% increase in total signal intensity found in the 60–110 p.p.m. region in progressing from plankton to deep trap material at the equatorial Pacific site. Most compositional changes noted both with depth and between sites were <5% of total signal intensity. Predominant NMR signals correspond to nominal 'alkyl' (0–45 p.p.m.), 'amino' (45–60 p.p.m.) and 'carbonyl' (165–215 p.p.m.) carbon, with the last including primarily carbon double-bonded to oxygen in esters or amides resonating near 175 p.p.m.²⁰. Also evident are O-alkyl carbon forms, including aliphatic carbon atoms single-bonded to one (60–95 p.p.m.) or two (~95–110 p.p.m.) oxygen atoms. The former O-alkyl resonance is more intense in all samples, and is typical of pentoses and hexoses. Small amounts of unsaturated carbon (110–165 p.p.m.) can also be detected, although it is not possible to determine from these measurements alone the extent to which this C = C occurs in aliphatic versus aromatic systems. Although these spectra and corresponding area percentages within the chemical shift intervals (Table 1) indicate a general increase in O-alkyl carbon (30–50% of that present in the plankton samples), this trend has little effect on overall distributions of spectral area.

These NMR results can be recast as a mixture of nominal biochemical equivalents²¹ to compare with compositions measured by conventional methods for the Pacific⁴ and other marine samples. To estimate biochemical compositions, only 'amino acid', 'carbohydrate' and 'lipid' carbon are assumed to be present (see Methods). These biochemical categories are over-generalizations that do not specifically include less abundant biochemical types such as nucleic acids and pigments. The weight percentages of biochemicals calculated from the NMR data change little down both water columns (Fig. 2). 'Amino acid' accounts for the largest component (40–50%) of POC in all samples, with 'carbohydrate' and 'lipid' making up roughly 30 and 20% of the remaining mass. A general trend down the water column towards increasing 'carbohydrate' is observed. Essentially all unsaturated carbon can be assigned to aromatic amino acids, without including other sources or reaction products. Although generally marginal in quality due to limited sample amount, solid-state ¹⁵N NMR spectra (not shown) obtained for the plankton and upper sediment-trap samples from the equatorial Pacific site exhibited only one discernible resonance near 360 p.p.m., as is typical of amide nitrogen. This observation independently supports the assumptions that most of the nitrogen in these samples resides in proteins, rather than in 'pyrrole-like' heterocyclic forms such as occur in chlorophyll, RNA, DNA and melanoidin-like humification products²².

Table 1 Composition of equatorial Pacific and Arabian Sea samples

Sample	Depth (m)	OC (wt%)	C/N (atom)	Flux (% of PP)	0–45* (Area%)	45–60* (Area%)	60–110* (Area%)	110–165* (Area%)	165–215* (Area%)
Equatorial Pacific									
Plankton	0–100	22.4	6.98	100	44.9	14.0	20.8	6.0	14.3
Upper trap	955	8.87	7.08	1.0	39.1	11.8	24.3	10.0	14.8
Deep trap	3459	5.69	8.06	0.5	40.1	15.6	29.3	4.2	10.8
Arabian Sea									
Plankton	0–100	24.0	7.37	100	45.6	12.4	22.3	7.6	12.0
Upper trap	498	9.08	9.53	1.5	41.5	11.1	25.9	9.3	12.2
Deep trap	2896	6.10	9.68	1.0	38.7	11.0	27.0	11.4	11.8

Equatorial Pacific Ocean, 0° N, 140° W; Arabian Sea, 17° N, 60° W. OC, organic carbon (salt-corrected); flux, given as a percentage of the average net primary production (PP) measured at the sampling time.

* Signal intensity within the indicated chemical shift intervals (in parts per million) expressed as a percentage of total signal intensity between 0 and 215 p.p.m.

In spite of the above simplifications, the modelled NMR results (Fig. 2) indicate major biochemical compositions that are similar to those obtained by direct chromatographic analysis of these same Pacific plankton and sediment-trap samples⁴. This correspondence holds even for the deep trap sample in which only ~20% of the POC can be chromatographically measured. The strongest lines of evidence that the NMR-based biochemical calculations in Fig. 2 are reasonable come from the observations that these compositions (1) agree closely with those directly measured for equatorial Pacific plankton in which >80% of carbon is chromatographically represented⁶, and also (2) fall within the range of literature values determined chemically for phytoplankton from various cultures and field studies^{23–25}.

We find no consistent evidence from the more comprehensive NMR data (Table 1) for either an increase in alkyl carbon corresponding to selective concentration of aliphatic algaenan¹³ biopolymers, or for an increase of unsaturated carbon as is typical of humification²⁶. The observed constancy is notable because the >98% attenuation in sinking POC flux (Table 1) is sufficient to concentrate a trace (<1%) initial component, or an intermediate reaction product, into a major constituent of the deep-water samples. It must follow by mass balance that most of the lost organic matter was recycled in the water column with little or no

chemical selectivity. This commonality is consistent with the observations that profiles of raining POC attenuate similarly between water depths of 100 m and thousands of metres at both study sites¹⁰, with minimal attending change in C/N ratios¹⁵. Whether organic degradation is similarly nonselective in marine sediments, or at sites (or times) characterized by different biological communities, remains to be seen.

Although these results provide strong evidence against substantial selective degradation and humification as sinking ocean particles are extensively remineralized, other mechanisms might lead to biochemical uniformity. For example, subtle changes in the stereochemistry and crosslinking of proteins and polysaccharides comprising plankton might impart increased resistance to biodegradation¹¹ and yet not be detectable by solid-state ¹³C NMR. Conversely, intrinsically resistant biomacromolecules exhibiting a bulk chemical composition resembling plankton might persist to comprise most particulate organic matter at depth. One natural product matching this criterion is peptidoglycan, a “chain-mail-like” macromolecule in bacterial cell walls that comprises roughly equal amounts of amino sugar strands with peptide bridges¹³. Downward increases of glycine²⁷ and total carbohydrate (Fig. 2) in these samples are consistent with a growing peptidoglycan content, but are inconsistent with most other evidence for a largely algal origin^{4,15,27}.

Given that inorganic matter makes up more than 80% of the total mass of all the sinking particle mixtures (Table 1), it seems more likely that the observed composition similarity results from physical protection of the organic fraction by predominant mineral components, such as opal, calcium carbonate and detrital aluminosilicates¹⁵. Physical shielding of soil and sedimentary organic substances by inorganic matter is widely documented^{21,28}, and could protect molecules within or between mineral grains. In addition, mineral grains exhibit densities relative to sea water that are at least an order of magnitude greater than those of pure organic substances, and hence can act as ballast⁵, proportionately increasing the sinking rate of any associated biochemicals^{29,30}. Individually, or in combination, mineral protection and ballasting of organic matter could strongly affect penetration of bioactive elements into the inner ocean, transfer of nutrients to benthic organisms and eventual preservation of organic matter in marine sediments. A predictive understanding of the ocean’s biological pump may require increased attention to how the organic and mineral components of the raining particles are joined, and in what ways these associations affect the comparative fates of both moieties¹⁰. □

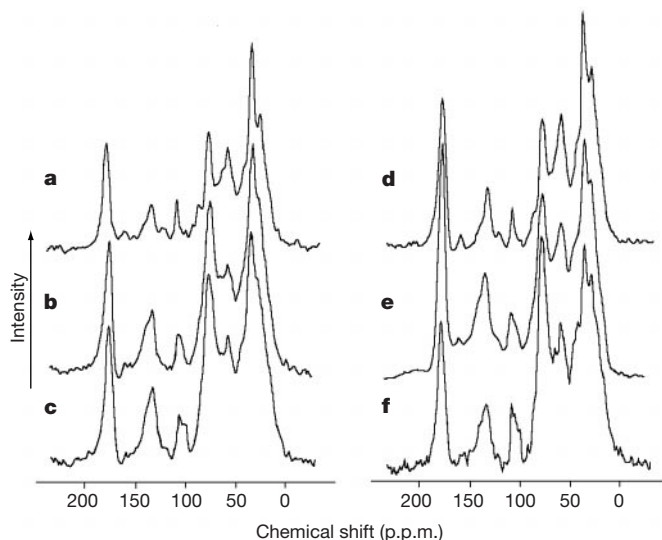


Figure 1 Solid-state ¹³C NMR spectra of material from the equatorial Pacific Ocean and the Arabian Sea. **a–c**, Material from the equatorial Pacific Ocean; **a**, plankton, **b**, upper trap, and **c**, deep trap samples. **d–f**, Material from the Arabian Sea; **d**, plankton, **e**, upper trap, and **f**, deep trap samples (Table 1).

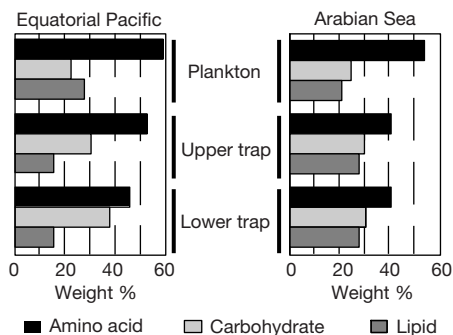


Figure 2 Calculated weight percentages of biochemicals in the samples from the equatorial Pacific Ocean and the Arabian Sea. See Methods for a description of the techniques used to acquire and model these spectral data.

Methods

Solid-state ¹³C NMR spectra were acquired at a ¹³C frequency of 50.3 MHz on a Varian Unity 200 spectrometer. Samples were packed into a 7-mm-diameter cylindrical zirconia rotor with Kel-F end caps, and spun at 5,000 ± 100 Hz in a Doty Scientific MAS (magic-angle spinning) probe. A conventional cross-polarization pulse sequence was used with a 1.0 ms contact time. An inversion recovery pulse sequence was employed to estimate relaxation times (T₁H values) for each type of sample. These relaxation times were used to set the duration of the recycle delay between pulses (0.3 s) to ≥7 times the longest measured T₁H value⁵. The number of transients collected for individual samples ranged between 22,000 and 262,000, depending on the amount of carbon that could be placed in the rotor. All spectra were processed with 50 Hz lorentzian line broadening and 0.010 s gaussian broadening. Spectral distributions for each sample were determined by integrating signal intensities in the five chemical-shift regions given in Table 1. Chemical shifts were externally referenced to the methyl resonance of hexamethylbenzene at 17.36 p.p.m. Signal intensities associated with spinning side bands of all major resonances above 110 p.p.m. consistently accounted for <1% of the parent peak intensity, and were numerically allocated back to the spectral region from which they derived.

Relative intensities expected for a typical marine amino-acid mixture were calculated on the basis of average mole percentages of individual amino acids in marine phytoplankton³¹. The carbon atoms in each amino acid were assigned proportionately to one of the five previously mentioned categories—alkyl, amino, O-alkyl, C = C and carboxyl. The ‘carbohydrate’ component was assumed to be a hexose polymer. The ‘lipid’ component was assigned the structure of stearic acid, because fatty acids are major components of lipids in phytoplankton and sinking marine particles⁴ and other lipids and algaenans also contain small amounts of oxygen¹⁵. This very basic mixing model allotted all nitrogen to ‘amino acid’. Carbons bonded to oxygen or nitrogen in lipids were assigned

to corresponding 'protein' and 'carbohydrate' components. The fraction of protein carbon in each spectral region was estimated on the basis of total nitrogen content²³ and the distribution of carbon types in the typical marine amino-acid mixture. The signal intensity remaining in the 0–45 p.p.m. and 60–110 p.p.m. regions after accounting for protein carbon was assigned to 'lipid' and 'carbohydrate' carbon, respectively. Calculated percentages of carbon in the three biochemical endmembers were then converted to weight percentages in a corresponding hydrolysate mixture by dividing by the fraction of carbon in each compound type and adding water to individual amino acids and sugars.

Received 3 April; accepted 15 November 2000.

1. Broecker, W. S. & Peng, T.-H. *Tracers in the Sea* (Lamont-Doherty, Palisades, New York, 1982).
2. Siegenthaler, U. & Sarmiento, J. L. Atmospheric carbon dioxide and the ocean. *Nature* **365**, 119–125 (1993).
3. Sarmiento, J. L. & Toggweiler, J. R. A new model for the role of the oceans in determining atmospheric pCO₂. *Nature* **308**, 621–624 (1984).
4. Wakeham, S. G., Lee, C., Hedges, J. I., Hernes, P. J. & Peterson, M. L. Molecular indicators of diagenetic wilson in marine organic matter. *Geochim. Cosmochim. Acta* **61**, 5363–5369 (1997).
5. Wilson, M. A. *NMR Techniques and Applications in Geochemistry and Soil Chemistry* (Pergamon, Oxford, 1987).
6. Lee, C. *et al.* Particulate organic carbon fluxes: compilation of results from the 1995 US JGOFS Arabian Sea Process Study. *Deep-Sea Res.* **45**, 2489–2501 (1998).
7. Stevenson, F. J. *Humus Chemistry* (Wiley, New York, 1994).
8. Hatcher, P. G., Spiker, E. C., Szeverenyi, N. M. & Maciel, G. E. Selective preservation and origin of petroleum-forming aquatic kerogen. *Nature* **305**, 498–501 (1983).
9. Martin, J. H., Knauer, G. A., Karl, D. M. & Broenkow, W. W. VERTEX: carbon cycling in the northeast Pacific. *Deep-Sea Res.* **34**, 267–285 (1987).
10. Armstrong, R. A., Lee, C., Hedges, J. I., Honjo, S. & Wakeham, S. G. A new model for deep-ocean remineralization of organic carbon and mineral ballasts. *Deep-Sea Res.* (in the press).
11. Hedges, J. I. *et al.* The molecularly uncharacterized component of nonliving organic matter in natural environments. *Org. Geochem.* **31**, 945–958 (2000).
12. Boyd, P. W. & Newton, P. P. Does plankton community structure determine downward particulate organic carbon flux in different oceanic provinces? *Deep-Sea Res.* **46**, 63–91 (1999).
13. De Leeuw, J. W. & Largeau, C. in *Organic Geochemistry* (eds Engel, M. & Macko, S. A.) 23–72 (Plenum, New York, 1993).
14. Knicker, H., Scaroni, A. W. & Hatcher, P. G. ¹³C and ¹⁵N NMR spectroscopic investigation on the formation of fossil algal residues. *Org. Geochem.* **24**, 661–669 (1996).
15. Hernes, P. J. *et al.* Particulate carbon and nitrogen fluxes and compositions in the central Equatorial Pacific. *Deep-Sea Res.* (in the press).
16. Murray, J. W., Barber, R. T., Roman, M. R., Bacon, M. P. & Feely, R. A. Physical and biological controls on carbon cycling in the equatorial Pacific. *Science* **266**, 58–65 (1994).
17. Smith, S. L., Codispoti, L. A. & Morrison, J. M. The US/JGOFS process study in the Arabian Sea: Introduction. *Deep-Sea Res.* **45**, 2053–2101 (1998).
18. Honjo, S., Dymond, J., Prell, W. & Ittekkot, V. Monsoon-controlled export fluxes to the interior of the Arabian Sea. *Deep-Sea Res.* **46**, 1859–1902 (1999).
19. Kinches, P., Powlson, D. S. & Randall, E. W. ¹³C NMR studies of organic matter in whole soils: 1. Quantitation possibilities. *Eur. J. Soil Sci.* **46**, 125–138 (1995).
20. Baldock, J. A. & Nelson, P. N. in *Handbook of Soil Science* (eds Sumner, M. E. *et al.*) B25–B84 (CRC, Boca Raton, 2000).
21. Nelson, P. N., Baldock, J. A., Oades, J. M. & Churchman, G. J. Dispersed clay and organic matter in soil: their nature and associations. *Aust. J. Soil Res.* **37**, 289–315 (1999).
22. McCarthy, M., Pratum, T., Hedges, J. & Benner, R. Chemical composition of dissolved organic nitrogen in the ocean. *Nature* **390**, 150–154 (1997).
23. Laws, E. A. Photosynthetic quotients, new production and net community production in the open ocean. *Deep-Sea Res.* **38**, 143–167 (1991).
24. Anderson, L. A. On the hydrogen and oxygen content of marine phytoplankton. *Deep-Sea Res.* **38**, 143–167 (1995).
25. Parsons, T. R., Stephens, K. & Strickland, J. D. H. On the chemical composition of eleven species of marine phytoplankters. *J. Fish. Res. Bd. Can.* **18**, 1001–1015 (1961).
26. Hedges, J. I. in *Humic Substances and Their Role in the Environment* (eds Frimmel, F. H. & Christman, R. F.) 45–58 (Dahlem Konferenzen, West Berlin, 1988).
27. Lee, C., Wakeham, S. G. & Hedges, J. I. Composition and flux of particulate amino acids and pigments in equatorial Pacific seawater and sediments. *Deep-Sea Res.* **47**, 1535–1568 (2000).
28. Hedges, J. I. & Oades, J. M. Comparative organic geochemistries of soils and marine sediments. *Org. Geochem.* **27**, 319–361 (1997).
29. Bishop, J. K. B., Edmond, J. M., Ketten, D. R., Bacon, M. P. & Silker, W. B. The chemistry, biology, and vertical flux of particulate matter from the upper 400m of the equatorial Atlantic Ocean. *Deep-Sea Res.* **24**, 511–548 (1977).
30. Ittekkot, V., Haake, B., Bartsch, M., Nair, R. R. & Ramaswamy, V. in *Upwelling Systems: Evolution since the Miocene* (eds Summerhayes, C. P., Prell, W. L. & Emeis, K. C.) 167–176 (Special Publication No. 64, Geological Society, London, 1992).
31. Cowie, G. L. & Hedges, J. I. Sources and reactivities of amino acids in a coastal marine environment. *Limnol. Oceanogr.* **37**, 703–724 (1992).

Acknowledgements

We thank P. Hernes, B. Bergamaschi, J. Murray, J. Dymond and the captain and crew of the RV *Thomas G. Thompson* for support during the cruises; S. Honjo and J. Dymond (and field groups) for accommodating our sediment traps on their moorings; and A. Aufdenkampe, R. Benner and B. van Mooy for comments that significantly improved the manuscript. This work was supported by the NSF (J.H., C.L. and S.W.); Y.G. thanks the Canadian Natural Sciences and Engineering Research Council (NSERC) for a post-doctoral fellowship.

Correspondence and requests for materials should be addressed to J.H. (e-mail: jihedges@u.washington.edu).

Interhemispheric climate links revealed by a late-glacial cooling episode in southern Chile

Patricio I. Moreno^{*†}, George L. Jacobson Jr^{*‡}, Thomas V. Lowell[§] & George H. Denton^{*||}

^{*} Institute for Quaternary and Climate Studies, [‡] Department of Biological Sciences, ^{||} Department of Geological Sciences, University of Maine, Orono, Maine 04469, USA

[§] Department of Geology, University of Cincinnati, Cincinnati, Ohio 45221, USA

Understanding the relative timings of climate events in the Northern and Southern hemispheres is a prerequisite for determining the causes of abrupt climate changes. But climate records from the Patagonian Andes^{1–4} and New Zealand^{5–8} for the period of transition from glacial to interglacial conditions—about 14.6–10 kyr before present, as determined by radiocarbon dating—show varying degrees of correlation with similar records from the Northern Hemisphere. It is necessary to resolve these apparent discrepancies in order to be able to assess the relative roles of Northern Hemisphere ice sheets and oceanic, atmospheric and astronomical influences in initiating climate change in the late-glacial period. Here we report pollen records from three sites in the Lake District of southern Chile (41°S) from which we infer conditions similar to modern climate between about 13 and 12.2 ¹⁴Ckyr before present (BP), followed by cooling events at about 12.2 and 11.4 ¹⁴Ckyr BP, and then by a warming at about 9.8 ¹⁴Ckyr BP. These events were nearly synchronous with important palaeoclimate changes recorded in the North Atlantic region⁹, supporting the idea that interhemispheric linkage through the atmosphere was the primary control on climate during the last deglaciation. In other regions of the Southern Hemisphere, where climate events are not in phase with those in the Northern Hemisphere, local oceanic influences may have counteracted the effects that propagated through the atmosphere.

Southern Chile is ideal for the study of interhemispheric linkages throughout the Quaternary period for three reasons: (1) it has insolation regimes that are out-of-phase with northern mid-latitudes, (2) it is located on the windward side of southern South America, making it highly sensitive to variations in the southern westerlies, and (3) it is far removed from the direct influence of Northern Hemisphere ice sheets and sites of North Atlantic Ocean thermohaline convection. The southern westerlies carry the precipitation that nourishes alpine glaciers and temperate rainforests in southern Chile. Modern vegetation in the Chilean Lake District (38–42°S) includes—from low to high altitudes—the Valdivian, North Patagonian (NPRF) and subantarctic rainforest communities¹⁰. The chief components of the low- to mid-elevation Valdivian rainforest community are *Eucryphia*, *Nothofagus dombeii*, *Nothofagus obliqua*, several species of Myrtaceae and Proteaceae, epiphytic ferns, and woody lianas. Cupressaceae and Podocarpaceae are characteristic of the NPRF community above 500 m elevation^{11,12}. At 41°S, the deciduous *Nothofagus pumilio* and *Nothofagus antarctica* dominate the subantarctic rainforests near to the tree line at ~1,100 m elevation, along with composite and ericaceous shrubs, and herbs.

We developed pollen records from three sites in the lowlands of the Lake District (Fig. 1). Glacial deposits and landforms corresponding to the last ice age have been examined in detail in this

[†] Present address: Departamento de Biología, Facultad de Ciencias, Universidad de Chile, Casilla 653, Santiago, Chile.

region^{1,2,12–14}; these studies show an abrupt withdrawal of Andean piedmont glacier lobes at about 14.6 ¹⁴C kyr BP, marking the onset of the last termination.

Close-interval sampling and dating of sediment cores (Table 1) from the Canal de la Puntilla (40° 56' S, 72° 54' W, 120 m above sea level), Huelmo (41° 31' S, 73° 00' W, 25 m.a.s.l.) and Lago Condorito sites (41° 45' S, 73° 07' W, 100 m.a.s.l.) (Fig. 1) yielded pollen records with an average time resolution of ~55 ¹⁴C years between samples. These sites show a coherent pollen assemblage between

~13 and 12.2 ¹⁴C kyr BP with abundant trees and vines, along with forest ferns (Figs 2, 3). This assemblage indicates the predominance of closed-canopy NPRF under a temperate and humid climate, suggesting conditions approaching modern climate in the foothills of the mountain ranges at ~41° S. The expansion of *Podocarpus* at ~12.2 ¹⁴C kyr BP and the persistence of rainforest vegetation (Figs 2, 3) are consistent with an onset of cooler conditions, as suggested by the latitudinal and altitudinal distribution of conifers in modern NPRF^{10,11}. This vegetation change coincided with a prominent

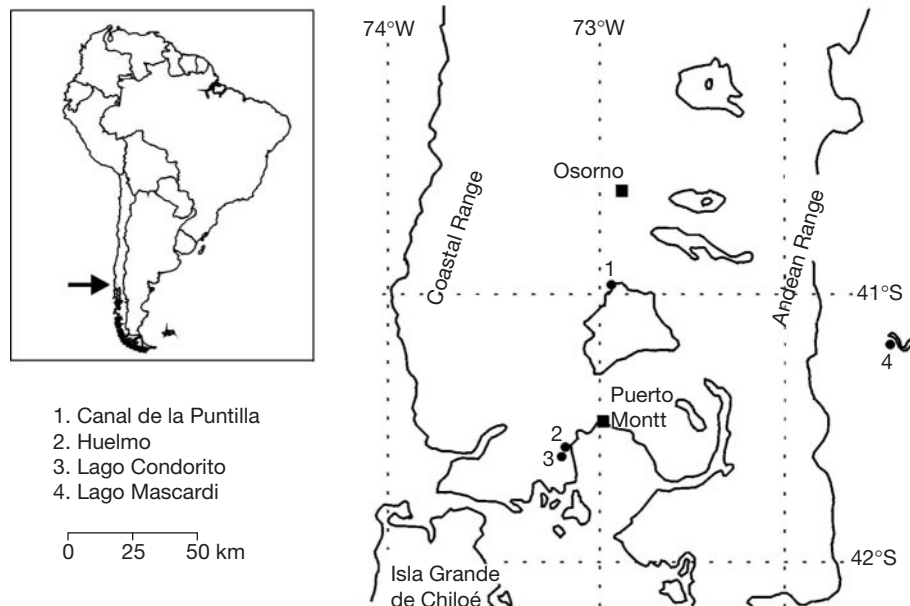


Figure 1 Location map, showing the pollen sites in the Chilean Lake District and the Lago Mascardi site in Argentina. All sediment cores were obtained using a 5-cm-diameter Wright piston corer. We selected small, closed-basin sites to obtain our sediment cores in order to establish comparisons among sites having similar depositional environments. Sites differ in hydrology as only one of them has persisted as a lake (the Lago Condorito site), while the other sites have undergone terrestrialization processes since the mid-Holocene (Huelmo), or have completely dried out during the early Holocene (the Canal de la Puntilla site). These contrasts in hydrology, along with local differences in microclimate

and topography, account for the differences in present vegetation and past pollen assemblages. Despite these differences, all records show the same timing, character, and direction of vegetation change (Figs 2, 3). There is a gradient of decreasing summer and total annual precipitation towards the north, due to a lower frequency of storm tracks, intensification of the rain-shadow effect caused by the Coastal Range, which increases in altitude toward the north, and the maritime influence of the Seno Reloncaví seaway in the south. Precipitation occurs evenly throughout the year in the Seno Reloncaví sector of the Lake District, with decreased frontal activity during the summer months.

Table 1 Radiocarbon dates and $\delta^{13}\text{C}$ values

Site	Depth (cm)	Material	¹⁴ C yr BP	$\delta^{13}\text{C}$ (‰)	Laboratory no.*
Canal de la Puntilla	108–109	Bulk	9,820 ± 70	–28.9	AA-11079
Canal de la Puntilla	111–112	Wood	10,040 ± 80	–31	ETH-17640
Canal de la Puntilla	122–124	Charcoal	10,430 ± 100	–21.7	ETH-17641
Canal de la Puntilla	137–139	Bulk	10,665 ± 120	–29.6	A-7755
Canal de la Puntilla	148–151	Bulk	12,180 ± 120	–29.6	A-7756
Canal de la Puntilla	158–160	Bulk	12,620 ± 140	–29.6	A-7757
Canal de la Puntilla	178–180	Bulk	13,290 ± 140	–29.6	A-7758
Huelmo	704–708	Wood	9,635 ± 65	–30.3	AA-23236
Huelmo	752–756	Bulk	10,585 ± 70	–33.2	AA-23237
Huelmo	784–788	Bulk	11,460 ± 130	–31.7	AA-23238
Huelmo	812–816	Wood	11,930 ± 75	–29.6	AA-23239
Huelmo	888–892	Bulk	12,825 ± 80	–32.7	AA-23240
Lago Condorito	613–615	Bulk	9,110 ± 45	–30.1	A-8587
Lago Condorito	685–688	Bulk	9,680 ± 85	–30.5	A-8070
Lago Condorito	724–726	Bulk	10,060 ± 60	–28.4	A-8069
Lago Condorito	821–824	Bulk	11,265 ± 65	–29.9	A-8068
Lago Condorito	848–852	Bulk	11,435 ± 80	–29.5	A-8067
Lago Condorito	863–873	Bulk	12,230 ± 140	–30	Beta-60352
Lago Condorito	873–886	Bulk	12,170 ± 90	–32.1	A-8066
Lago Condorito	928–930	Bulk	12,330 ± 130	–29	A-6661

The lowlands of the Chilean Lake District are filled with Quaternary sediments, comprised mainly of glacial, fluvio-glacial and volcanic deposits. The absence of carbonate bedrock near the study sites rules out the possibility of contamination of our radiocarbon dates with ambient old carbon. In addition, only small lakes with shallow and unconfined aquifers with short residence times were sampled. Under these conditions, the reservoir effect on these small water bodies is negligible. We have obtained more than 650 radiocarbon dates as part of our joint glacial geologic and palynological investigations in the Lake District, and have found no significant differences between conventional radiocarbon dates on bulk organic sediment or terrestrial macrofossil samples, and accelerator mass spectrometry (AMS) dates on bulk organic sediment or terrestrial macrofossils. The PDB standard was used in determining $\delta^{13}\text{C}$ values.

* Abbreviations: AA, the NSF-Arizona AMS facility; ETH, the Swiss Federal Institute of Technology AMS Facility; A, University of Arizona Geochemistry Laboratory; and Beta, Beta Analytic.

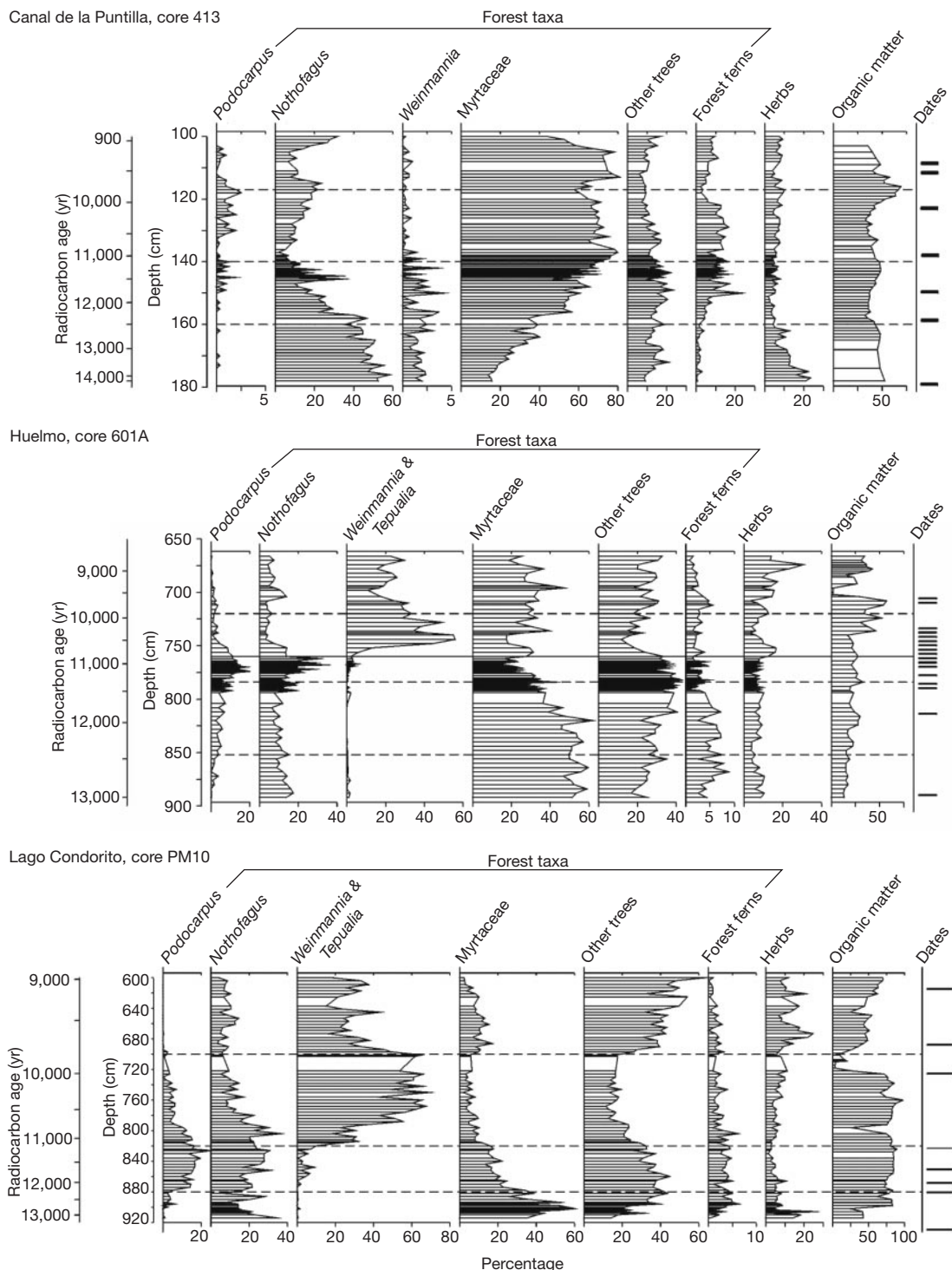


Figure 2 Percentage diagrams of selected taxa from the Canal de la Puntilla, Huelmo and Lago Condorito sites. Percentage scales among species vary for visual depiction only. The horizontal dashed lines mark the onset of vegetation and climate changes at ~ 12.2 , ~ 11.4 and 9.8 ^{14}C kyr BP. The bars on the right-hand side of the diagrams indicate the mean depth of the radiocarbon dates listed in Table 1. We used polynomial equations to develop age models, in order to assign interpolated radiocarbon ages for the pollen, charcoal, and loss-on-ignition results. All age models yielded $R^2 > 0.95$ and $P < 0.001$. The abundance of *Nothofagus* shown in the pollen diagrams refers to the palynomorph *Nothofagus dombeyi*-type, which includes several species that occur in several forest formations in the temperate region of Chile and Argentina. Although these species occur along a broad latitudinal and altitudinal range, their relative contribution to the local pollen rain is diminished in the low-altitude, thermophilous forest communities. Thus, in the context of the pollen assemblages of Canal de la Puntilla, Huelmo and Lago Condorito between 13 and 9 ^{14}C kyr BP, the expansion of *Nothofagus* along with *Podocarpus*

nubigena at 12.2 and ~ 11.2 ^{14}C kyr strongly suggests the onset of cooler conditions. The modest increase in the percentages of *P. nubigena*, and the absence of *Pseudopanax laetevirens* and *Tepualia stipularis* in Canal de la Puntilla, suggests that a precipitation gradient between the northern and southern portions of the Lake District was in effect during late-glacial time. All sites were small lakes or ponds between ~ 13 and 9.8 ^{14}C kyr BP, as suggested by their notably similar and homogeneous lithologies (silty gyttja, organic silt, and a horizontally lain volcanic ash layer deposited at ~ 9.8 ^{14}C kyr BP). The only exception is Lago Condorito, which shows a transition from basal peat to gyttja at ~ 12.8 ^{14}C kyr BP. Undisturbed, continuous, *in situ* sediment deposition at all sites during late-glacial time is shown by multiple cores obtained close to each other at all sites, and is further supported by several lines of evidence (sediment and X-ray stratigraphy, magnetic susceptibility logs, loss-on-ignition analyses, presence of volcanic ash horizons, age-depth curves, and so on; data not shown).

transition from basal peat to gyttja in the Lago Condorito record, suggesting that the cooling event was accompanied by an increase in precipitation/effective moisture.

The Canal de la Puntilla and Huelmo sites show a decline in thermophilous taxa and expansion of *Podocarpus* and *Nothofagus* starting at ~ 11.4 ^{14}C kyr BP (Figs 2, 3). These cold-resistant trees persisted until ~ 9.8 ^{14}C kyr BP in all sites. We interpret these results as an intensification of the vegetation changes that started at ~ 12.2 ^{14}C kyr BP, thus suggesting another cooling event at ~ 11.4 ^{14}C kyr BP. Hygrophilous NPRF taxa persisted between ~ 11.4 and 9.8 ^{14}C kyr BP (arboreal pollen $>90\%$), suggesting conditions similar to the climate that now exists at mid-to-high elevations in the mountain ranges at $\sim 41^\circ\text{S}$. Subsequent warming at 9.8 ^{14}C kyr BP led to the expansion of thermophilous taxa and the decline/disappearance of *Podocarpus* in all sites.

The Huelmo and Lago Condorito sites show the spread of *Weinmannia* and *Tepualia* starting at 10.9 and 11.2 ^{14}C kyr BP, respectively (Figs 2, 3). *Weinmannia* is now found in several rain-forest communities along a broad latitudinal and altitudinal range in the temperate region of southern Chile. It is a shade-intolerant emergent tree, described as a long-lived pioneer species tolerant of infertile and poorly drained soils^{15,16}. In addition, *Tepualia* thickets are known to expand following disturbance in temperate forest communities¹⁷. Considering the autoecology of these species, the pollen assemblages between 11.4 and 9.8 ^{14}C kyr BP indicate a shift from a closed-canopy rainforest community to a woodland that was dominated by species favoured by local disturbance. This vegetation change coincides with a prominent increase in charcoal particles at ~ 11 ^{14}C kyr BP (Fig. 3), implying local disturbance by fire. The pollen sites showing the earliest and most severe effects of fire are

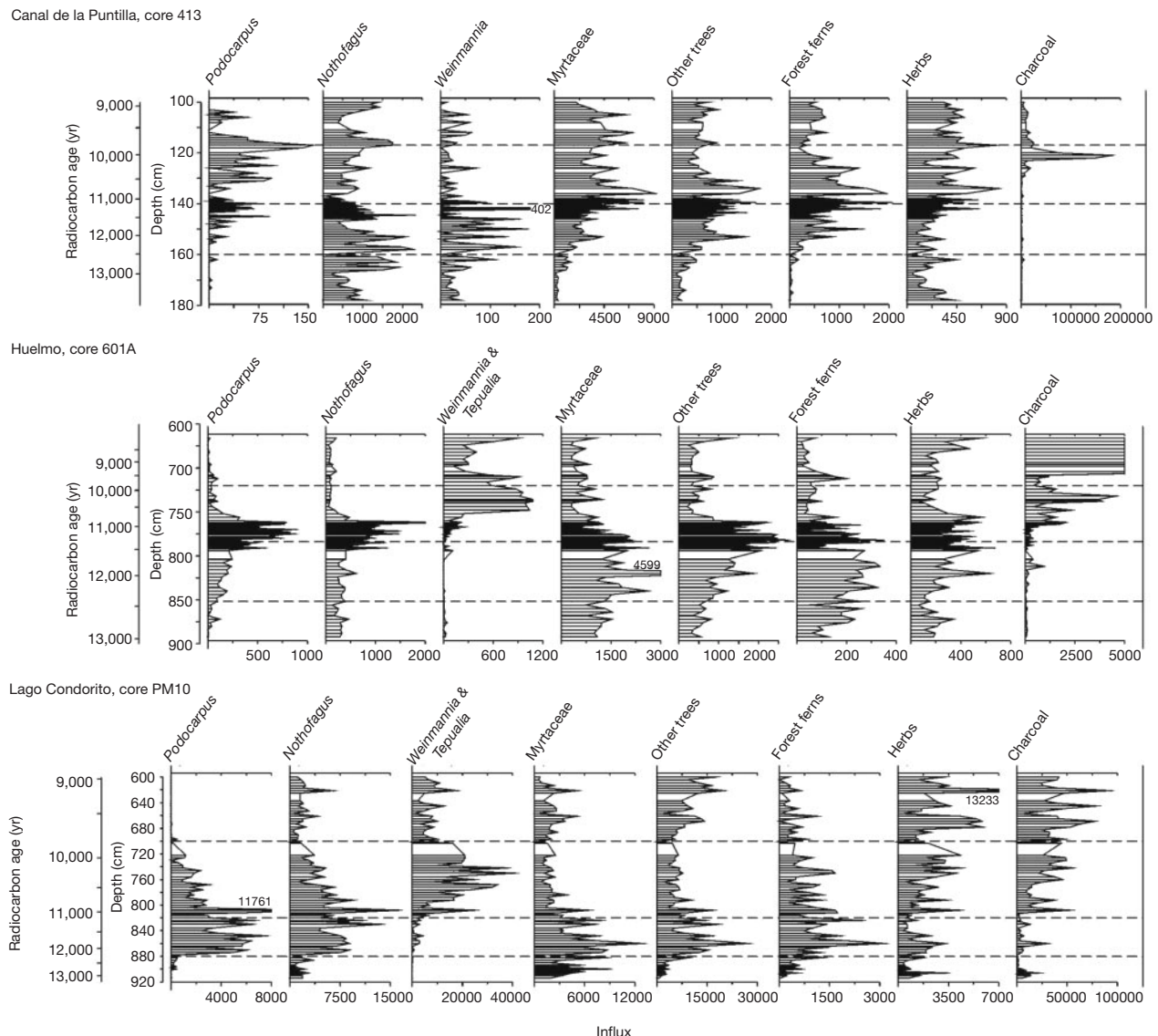


Figure 3 Influx diagrams of pollen, spores and charcoal particles from the Canal de la Puntilla, Huelmo and Lago Condorito sites. All sites show the expansion and persistence of *Podocarpus* between ~ 12.2 and 9.8 ^{14}C kyr BP. The Canal de la Puntilla and Huelmo records show a sharp increase of cold-resistant trees and declines in other trees and forest ferns at ~ 11.4 ^{14}C kyr BP. Unlike the Lago Condorito record, the onset of disturbance by fire in the Huelmo site (760 cm) lagged behind the cooling event at ~ 11.4 ^{14}C kyr BP (784 cm). The abrupt increase of charcoal in Canal de la Puntilla occurred $\sim 1,000$ ^{14}C years after the cooling event at ~ 11.4 ^{14}C kyr BP. An explanation

involving an accentuation of rainfall seasonality and/or continentality is inadequate to account for the occurrence of fires during late-glacial time, because the stratigraphic and geographical distribution of charcoal particles is inverse to the inferred precipitation gradient. Ignition by volcanic activity can be ruled out as a causal agent for the charcoal record, because of the lack of tephra layers between ~ 13 and 9.8 ^{14}C kyr BP at all sites. The influx scale for microscopic charcoal particles from Lago Condorito is truncated to emphasize the fire history from 13 to 9.5 ^{14}C kyr BP. Units of influx: grains or particles per cm^2 per yr.

located near Monte Verde, thought to be the oldest archaeological site in the Americas¹⁸. Remains interpreted as partially burned plant and animal artefacts in association with hearths have been cited as evidence for the presence of humans and their use of fire between ~12.5 and 12¹⁴Ckyr BP. Possibly, the high climate variability between ~11.2 and 9.9¹⁴Ckyr BP afforded the conditions that allowed humans to set small-scale fires that altered the vegetation near the Huelmo and Lago Condorito sites.

Our data suggest that climate approaching modern conditions prevailed in the Chilean Lake District between ~13 and 12.2¹⁴Ckyr BP. This was followed by a general reversal in trend with cooling events at ~12.2 and ~11.4¹⁴Ckyr BP, and then by subsequent warming at 9.8¹⁴Ckyr BP. The total temperature depression between ~11.4 and 9.8¹⁴Ckyr BP was relatively minor ($\leq 3^\circ\text{C}$), as indicated by the persistence of rainforest vegetation. The timing, direction, and relative magnitude of these events matches the late-glacial record from nearby Lago Mascardi¹⁹ (Fig. 1), which indicates retreat of the Monte Tronador ice cap between 13 and 12.4¹⁴Ckyr BP, a reversal starting at ~12.4¹⁴Ckyr BP and culminating with glacial readvance between 11.4 and 10.2¹⁴Ckyr BP. These records from the Andean region of mid-latitude South America show a notable resemblance in timing and structure to palaeoclimate fluctuations recorded in Europe and Greenland. In contrast, Bennett *et al.*⁴ found no palynological evidence for climate change during late-glacial time in the Chilean channels (45°–47°S). If this interpretation is correct, their results would imply that: (1) a major climate boundary existed between 42° and 45°S during late-glacial time; or (2) late-glacial climate changes did not reach a critical threshold to trigger discernible vegetation changes in palynological records, and thus the impoverished late-glacial flora of the Chilean channels was insensitive to the magnitude of late-glacial cooling; and/or (3) plant succession, soil development, and migration from glacial refugia were the dominant factors controlling vegetation change in this newly deglaciated region during the critical time period.

Our results suggest that mid-latitude climate in the Southern Hemisphere changed in unison with the North Atlantic region between ~13 and 10¹⁴Ckyr BP. This is in contrast with the palaeoclimate signal derived from sediment cores in the South Atlantic Ocean²⁰ and ice cores from interior Antarctica²¹, where the pattern of climate change is opposite to that found in the Chilean Lake District between ~13 and 10¹⁴Ckyr BP. Determining the representativeness of these opposing results on a hemispheric scale has important implications for understanding the climate mechanisms operative during ice ages, because one set of data supports an in-phase interhemispheric linkage in the atmosphere^{2,22,23}, whereas the other favours an out-of-phase relationship via a bipolar see-saw in deep ocean circulation²⁴. The solution could well be that synchronous climate changes, propagated in the atmosphere over much of the planet, were counteracted in Antarctica by a bipolar see-saw of thermohaline circulation, whose effects in the Southern Hemisphere were confined to the high southern latitudes. □

Received 12 February; accepted 28 November 2000.

- Denton, G. H. *et al.* Interhemispheric linkage of paleoclimate during the last glaciation. *Geogr. Ann.* **81**, 107–155 (1999).
- Lowell, T. V. *et al.* Interhemispheric correlation of Late Pleistocene glacial events. *Science* **269**, 1541–1549 (1995).
- Zhou, M. & Heusser, C. J. Late-glacial palynology of the Myrtaceae of Chile. *Rev. Palaeobot. Palynol.* **91**, 283–315 (1996).
- Bennett, K. D., Haberle, S. G. & Lumley, S. H. The last glacial-Holocene transition in Southern Chile. *Science* **290**, 325–328 (2000).
- Denton, G. H. & Hendy, C. H. The age of the Waiho Loop glacial event. *Science* **271**, 668–670 (1995).
- Denton, G. H. & Hendy, C. H. Younger Dryas age advance of Franz Josef Glacier in the Southern Alps of New Zealand. *Science* **264**, 1434–1437 (1994).
- Singer, C., Shulmeister, J. & McLea, B. Evidence against a significant Younger Dryas cooling event in New Zealand. *Science* **281**, 812–814 (1998).
- Newham, R. M. & Lowe, D. J. Fine-resolution pollen record of late-glacial climate reversal from New Zealand. *Geology* **28**, 759–762 (2000).

- Björk, S. *et al.* An event stratigraphy for the Last Termination in the North Atlantic region based on the Greenland ice-core record: a proposal by the INTIMATE group. *J. Quat. Sci.* **13**, 281–292 (1998).
- Villagrán, C. Vegetationsgeschichtliche und pflanzensoziologische Untersuchungen im Vicente Perez Rosales National park (Chile). *Diss. Bot.* **54**, 1–165 (1980).
- Páez, M. M., Villagrán, C. & Carrillo, R. Modelo de la dispersión polínica actual en la región templada chileno-argentina de Sudamérica y su relación con el clima y la vegetación. *Rev. Chil. Hist. Nat.* **67**, 417–434 (1994).
- Mercer, J. H. Glacial history of southernmost South America. *Quat. Res.* **6**, 125–166 (1976).
- Mercer, J. H. Chilean glacial chronology 20,000 to 11,000 carbon-14 years ago: some global comparisons. *Science* **172**, 1118–1120 (1972).
- Porter, S. C. Pleistocene glaciation in the southern lake district of Chile. *Quat. Res.* **16**, 263–292 (1981).
- Lusk, C. H. Gradient analysis and disturbance history of temperate forests of the coast range summit plateau, Valdivia, Chile. *Rev. Chil. Hist. Nat.* **69**, 401–411 (1996).
- Lusk, C. H. Long-lived light-demanding emergents in southern temperate forests: the case of *Weinmannia trichosperma* (Cunoniaceae) in southern Chile. *Plant Ecol.* **140**, 111–115 (1999).
- Ramírez, C. in *Monte Verde, a Late Pleistocene Settlement in Chile* Vol. I, *Paleoenvironment and Site Context* 53–87 (Smithsonian Institution Press, Washington, 1989).
- Dillehay, T. *Monte Verde, a Late Pleistocene Settlement in Chile*. Vol. II, *The Archaeological Context and Interpretation* (ed. Gould, R. A.) (Smithsonian Institution Press, Washington, 1997).
- Ariztegui, D., Bianchi, M. M., Masferrer, J., Lafargue, E. & Niessen, F. Interhemispheric synchrony of late-glacial climatic instability as recorded in proglacial Lake Mascardi, Argentina. *J. Quat. Sci.* **12**, 333–338 (1997).
- Charles, C. D., Lynch-Stieglitz, J., Ninnemann, U. S. & Fairbanks, R. G. Climate connections between the hemispheres revealed by deep sea sediment core/ice core correlations. *Earth Planet. Sci. Lett.* **142**, 19–27 (1996).
- Sowers, T. & Bender, M. Climate records covering the last deglaciation. *Science* **269**, 210–214 (1995).
- Broecker, W. S. & Denton, G. H. The role of ocean-atmosphere reorganizations in glacial cycles. *Quat. Sci. Rev.* **9**, 305–341 (1989).
- Broecker, W. S. Massive iceberg discharges as triggers for global climate change. *Nature* **372**, 421–424 (1994).
- Broecker, W. S. Paleocene circulation during the last deglaciation: A bipolar seesaw? *Paleoceanography* **13**, 119–121 (1998).

Acknowledgements

We thank I. Hajdas, W. Beck and T. Jull for their contribution to the development of the radiocarbon chronology of Canal de la Puntilla and Huelmo sites. This work was supported by the Office of Climate Dynamics of NSF, NOAA, the National Geographic Society, the Geological Society of America, the EPSCOR program of NSF, and a Fondecyt grant.

Correspondence and requests for materials should be addressed to P.I.M. (e-mail: pimoreno@uchile.cl).

Evidence of recent volcanic activity on the ultraslow-spreading Gakkel ridge

M. H. Edwards*, G. J. Kurras†, M. Tolstoy‡, D. R. Bohnenstiehl‡, B. J. Coakley§ & J. R. Cochran‡

*Hawaii Institute of Geophysics and Planetology, POST 815; †Department of Geology and Geophysics, POST 842A, School of Ocean and Earth Science and Technology, University of Hawaii at Manoa, 1680 East-West Road, Honolulu, Hawaii 96822, USA

‡Lamont-Doherty Earth Observatory, Columbia University, 61 Route 9W, Palisades, New York 10964, USA

§Department of Geology, Tulane University, New Orleans, Louisiana 70118, USA

Seafloor spreading is accommodated by volcanic and tectonic processes along the global mid-ocean ridge system. As spreading rate decreases the influence of volcanism also decreases^{1–4}, and it is unknown whether significant volcanism occurs at all at ultraslow spreading rates ($<1.5\text{ cm yr}^{-1}$). Here we present three-dimensional sonar maps of the Gakkel ridge, Earth's slowest-spreading mid-ocean ridge, located in the Arctic basin under the Arctic Ocean ice canopy. We acquired this data using hull-mounted sonars attached to a nuclear-powered submarine, the USS *Hawkbill*. Sidescan data for the ultraslow-spreading

($\sim 1.0 \text{ cm yr}^{-1}$) eastern Gakkel ridge depict two young volcanoes covering approximately 720 km^2 of an otherwise heavily sedimented axial valley. The western volcano coincides with the average location of epicentres for more than 250 teleseismic events detected^{5,26} in 1999, suggesting that an axial eruption was imaged shortly after its occurrence. These findings demonstrate that eruptions along the ultraslow-spreading Gakkel ridge are focused at discrete locations and appear to be more voluminous and occur more frequently than was previously thought.

The Arctic basin is the last of Earth's oceanic frontiers. Permanent pack ice covers the Arctic Ocean, restricting the free motion of surface ships, making it impossible to map the sea floor comprehensively. The ice canopy impedes the use of satellite altimetry data to derive the predicted bathymetry models that now exist for every other ocean^{6,7}. Nuclear submarines are ideal for Arctic research because they operate beneath the pack ice and are thus independent of surface conditions. In 1995 the US Navy and National Science Foundation cooperatively developed the Science Ice Exercises (SCICEX), a five-year programme supported by nuclear-powered Sturgeon-class submarines to study the ice canopy, oceanography, biology and geology of the Arctic basin. For SCICEX-98 and SCICEX-99 the US Navy's submarine USS *Hawkbill* was equipped with the Seafloor Characterization and Mapping Pods (SCAMP), a geophysical mapping system built to create the first three-dimensional maps of the Arctic sea floor. SCAMP instrumentation includes a 12-kHz Sidescan Swath Bathymetric Sonar (SSBS), a swept frequency (2.75 kHz to 6.75 kHz) High-Resolution Sub-bottom Profiler (HRSP), a BGM-3 gravimeter and the Data Acquisition and Quality Control System (DAQCS)⁸. We used the Submarine Inertial Navigation system to navigate under the ice, supplemented by occasional fixes from the Global Positioning

Satellite network when the USS *Hawkbill* surfaced. Our relative positional accuracy was better than 3 km.

The SCICEX-99 survey of Gakkel ridge was carried out at an operating depth of 225 m and a speed of 16 knots. The average usable swath width for bathymetry data is 10 km; the average swath width for sidescan data is 16 km. Typical sub-bottom penetration in sedimented areas is 100 m. SCICEX-98 and SCICEX-99 data provide approximately 100% bathymetric and sidescan coverage for the western Gakkel ridge rift valley and flanks out to 50 km on both sides of the ridge axis. Full spreading rates range from 1.33 cm yr^{-1} to 1.15 cm yr^{-1} (ref. 9) at the ends of this region (Fig. 1). During SCICEX-99 a reconnaissance survey of the eastern Gakkel ridge extended coverage of the axial zone out to 1.0 cm yr^{-1} full-spreading rate⁹.

One goal of the SCICEX surveys was to resolve a debate regarding the nature of volcanism at the ultraslow-spreading Gakkel ridge. The presence of lineated magnetic anomalies^{10,11} over the entire ridge suggests that seafloor volcanism occurs. Three bathymetric profiles across the axis of Gakkel ridge near 15°E depict a central high with 200 m relief¹² on the axial valley floor that may be a constructional ridge analogous to those observed on the slow-spreading Mid-Atlantic Ridge^{13,14}. However, theoretical modelling predicts that melt production should be diminished, or even inhibited, at spreading rates less than 1.5 cm yr^{-1} (refs 1–4). Analysis of SCICEX-96 gravity data suggests that the Gakkel ridge crust is anomalously thin¹⁵, supporting a hypothesis of diminished volcanism. The paucity of high-resolution bathymetry and sidescan data for Gakkel ridge has prevented unequivocal identification of any volcanic features until now.

SCICEX-99 sidescan and bathymetry data for the reconnaissance survey of the eastern Gakkel ridge show two amorphously shaped

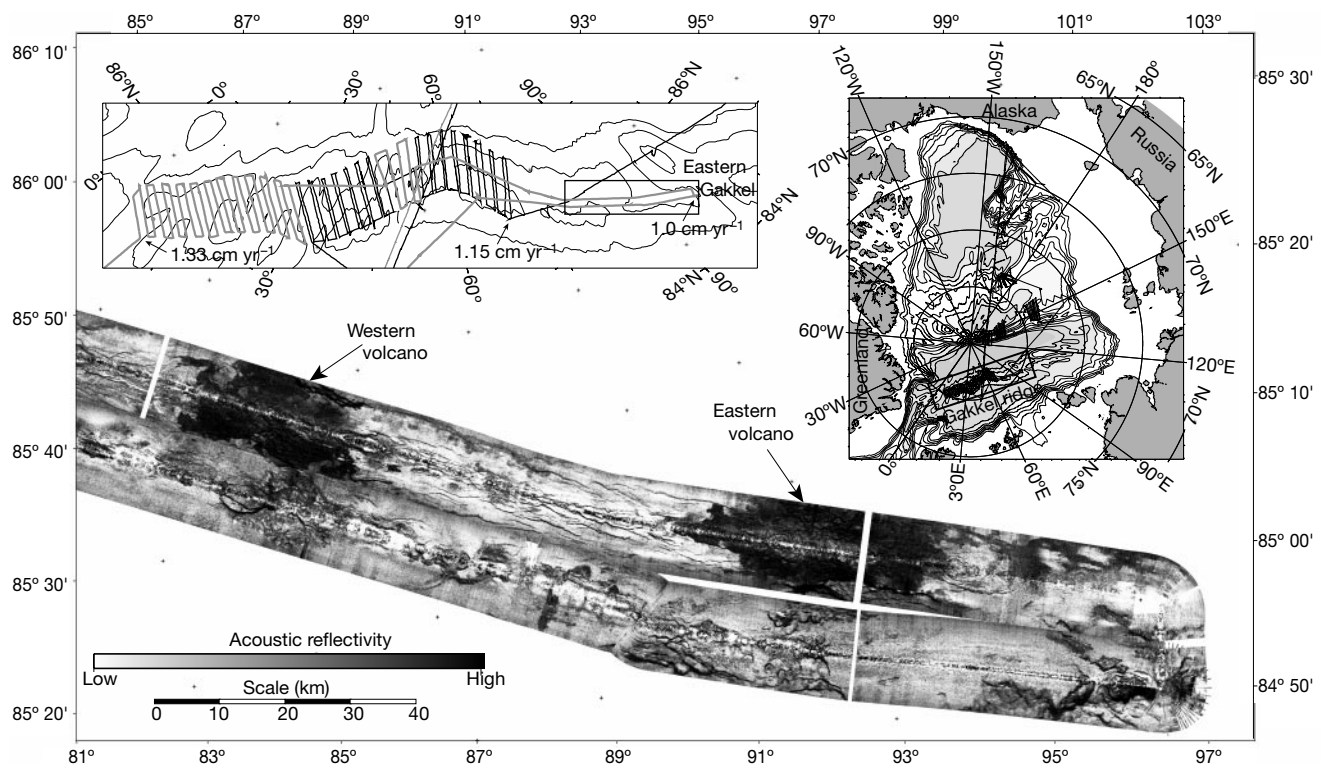


Figure 1 SCAMP sidescan data for the eastern Gakkel ridge. Two areas with high acoustic return strength, labelled the western volcano and eastern volcano, are probably younger than the weakly reflective terrain, presumably covered with thicker sediments, that surrounds them. Upper right inset, location map showing the Gakkel ridge survey within the Arctic basin. Bathymetric contour interval is 500 m; deeper regions are indicated by

lighter shades of grey. SCICEX-99 tracklines are shown in black. Upper left inset, detail of the SCICEX-98 (black) and SCICEX-99 (grey) tracklines for the Gakkel ridge surveys. Full spreading rates are indicated in three locations. The rectangle on the right side of the map shows the location of the sidescan data displayed below the insets.

areas with highly reflective acoustic character (Fig. 1) that are also topographic highs with 1,000 and 500 m relief, respectively (Figs 2 and 3). The sidescan data show long, sinuous channels of very reflective (dark grey) terrain, adjacent to and occasionally surrounding less reflective (light grey) terrain. The regions with lower acoustic reflectivity are interpreted to have thick sediment cover; the attenuation of sound waves in sediments reduces the strength of acoustic echoes¹⁶. Average sedimentation rates for the eastern Arctic are estimated to be 1–3 cm kyr⁻¹ (ref. 17), increasing as Gakkel ridge approaches the Laptev shelf. Given the 12.5-cm operational wavelength of the SSBS and the estimated sedimentation rate, it would take thousands of years to create sediment cover thick enough to attenuate the strength of the acoustic return significantly. We thus interpreted regions of low acoustic return to be older than strongly reflective regions.

Portions of the strongly reflective regions in the SCICEX-99 sidescan data abut lineaments, consistent with the ponding of lava against fault scarps (Fig. 2). Terminations of acoustically reflective terrain in regions devoid of lineaments have shapes characteristic of lava flow fronts or 'toes' (Fig. 3). The flow toes are radially distributed around the topographic highs. The morphology of the

strongly reflective regions is consistent with submarine volcanic flows mapped at other mid-ocean ridges by acoustic and optical systems^{18–20}. The acoustic character of the highly reflective Gakkel ridge terrain is very similar to a lava field at 8° S on the East Pacific Rise that was first detected because of its strong acoustic reflectivity¹⁸ and was subsequently shown to contain fresh, glassy basalts²¹. The two acoustically reflective regions are thus probably volcanoes that are largely devoid of sediment cover, and therefore erupted recently. The presence of these two young volcanoes, covering approximately 20% of the 3,750 km² surveyed along this portion of the eastern Gakkel ridge, proves that significant volcanism occurs at ultraslow spreading rates.

The sidescan data reveal that both volcanoes are cut by lineations interpreted to be faults formed by tectonic processes that occurred subsequent to volcanic emplacement. The western volcano (Fig. 3) is significantly less faulted than the eastern volcano (Fig. 2), suggesting that lava on the western volcano experienced less post-emplacement tectonism and is therefore younger. The few faults evident on the western volcano are located near the southern flank of the topographic high. Two of these faults traverse abrupt changes in acoustic reflectivity; these light/dark contacts indicate significant

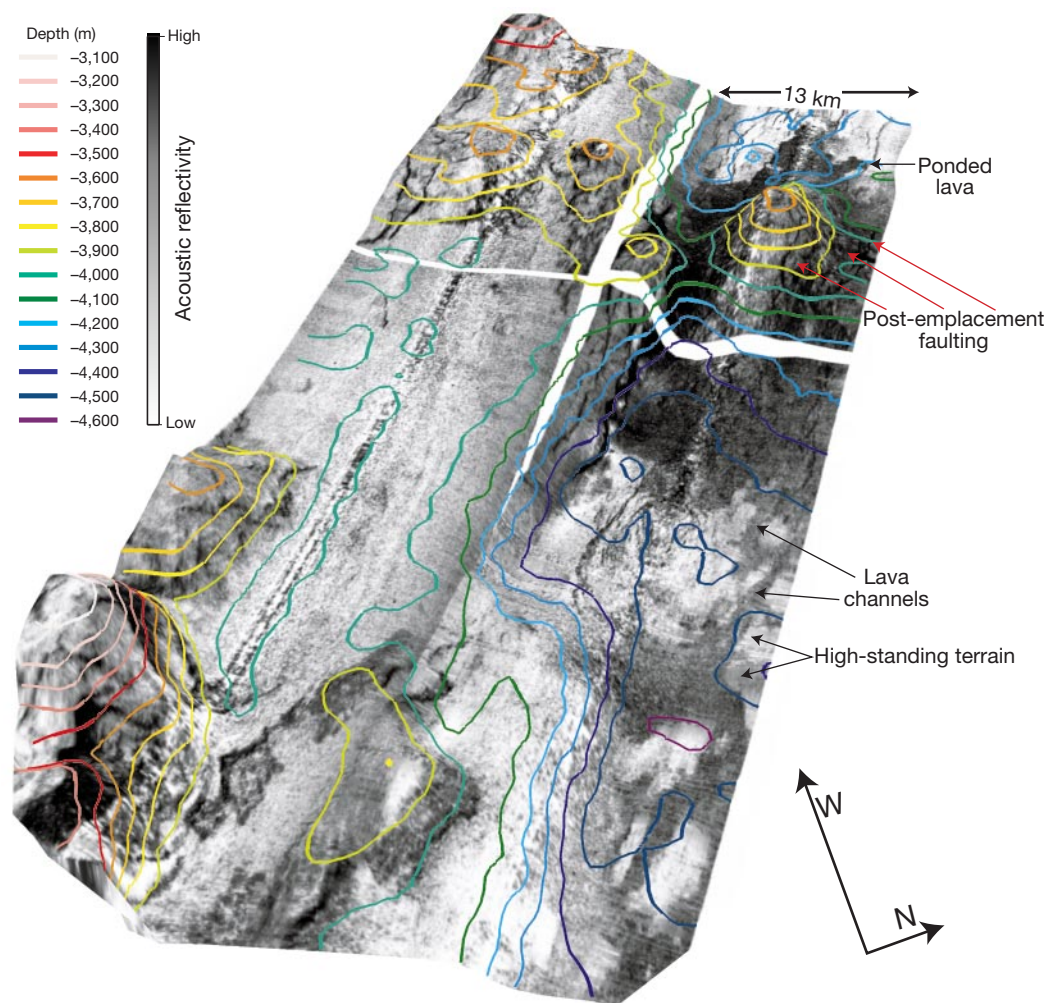


Figure 2 Three-dimensional view of the eastern volcano looking from east to west. In this image sidescan data are overlaid on a digital terrain model derived from SCAMP bathymetry. Colour-coded contours indicate depth. The region of highly reflective terrain located near the upper right corner of the data is approximately centred about a 1,000-m topographic high. To the west and east of the topographic high, sinuous channels of reflective terrain spill downslope. Where these channels adjoin lineations or higher-

elevation terrain they abruptly terminate or flow around features. This morphology is consistent with submarine lava flows observed on other mid-ocean ridges. The reflective terrain on the eastern volcano is laced with WNW–ESE trending lineations. These lineations are interpreted to be faults caused by tectonism that occurred after the emplacement of the reflective volcanic terrain. See Fig. 1 for the location of the eastern volcano.

differences in the amount of sediment cover. We interpret them to represent a boundary between lava flows of different ages. The presence of faults having sufficient vertical offset to be imaged by SCAMP indicates that the southern flank of the western volcano did not erupt recently. Assuming reasonable geological slip rates of a few mm yr^{-1} (refs 22, 23), faults of sufficient size to be imaged by SCAMP in $\sim 4,000$ m of water depth would require hundreds to several thousands of years to develop, although collapse events can form large scarps over short time periods²⁴. In contrast, the northern flank of the western volcano is remarkably devoid of lineations. In addition, all of the faults on the western volcano terminate abruptly to both the northwest and southeast. The abrupt truncation of faults on the western volcano and the absence of lineations in sidescan data for the northern flank support our hypothesis of a recent eruption that volcanically overprinted pre-existing faults. The volcanic overprinting and the presence of faults that cross-cut regions with different amounts of sediment cover lead us to conclude that the western volcano was formed by more than one eruption.

Teleseismic data for the Arctic basin corroborates the hypothesis

of a recent Gakkel ridge eruption. In January 1999 global seismic networks detected the beginning of an earthquake swarm on Gakkel ridge centred near 86°N , 85°E (refs 5, 26). Seismic activity continued through September 1999 although 75% of the events took place before the end of May. On 6 May 1999, the USS *Hawkbill* passed directly over the average location of the earthquake epicentres. The average location of the epicentres corresponds to the location of the western volcano (Fig. 3). The remarkable correlation between the locations of the earthquake epicentres and the location of the strongly reflective, un-tectonized western volcano together with the volcanic character of the seismic record^{5,26} provide evidence that lava erupted on the eastern Gakkel ridge days to months before SCAMP mapped the area. Because 12-kHz sonars can penetrate through thin sediments covering acoustically reflective lavas²⁵, it is possible that no eruption occurred on Gakkel ridge in 1999; however, historical global seismic records indicate that this is the only earthquake swarm detected on Gakkel ridge in about 100 years (ref. 5). SCAMP HRSP data show no evidence of sediment layering on either volcano although there is evidence of layering adjacent to both (Fig. 4). Taken together, the SCICEX and teleseismic data

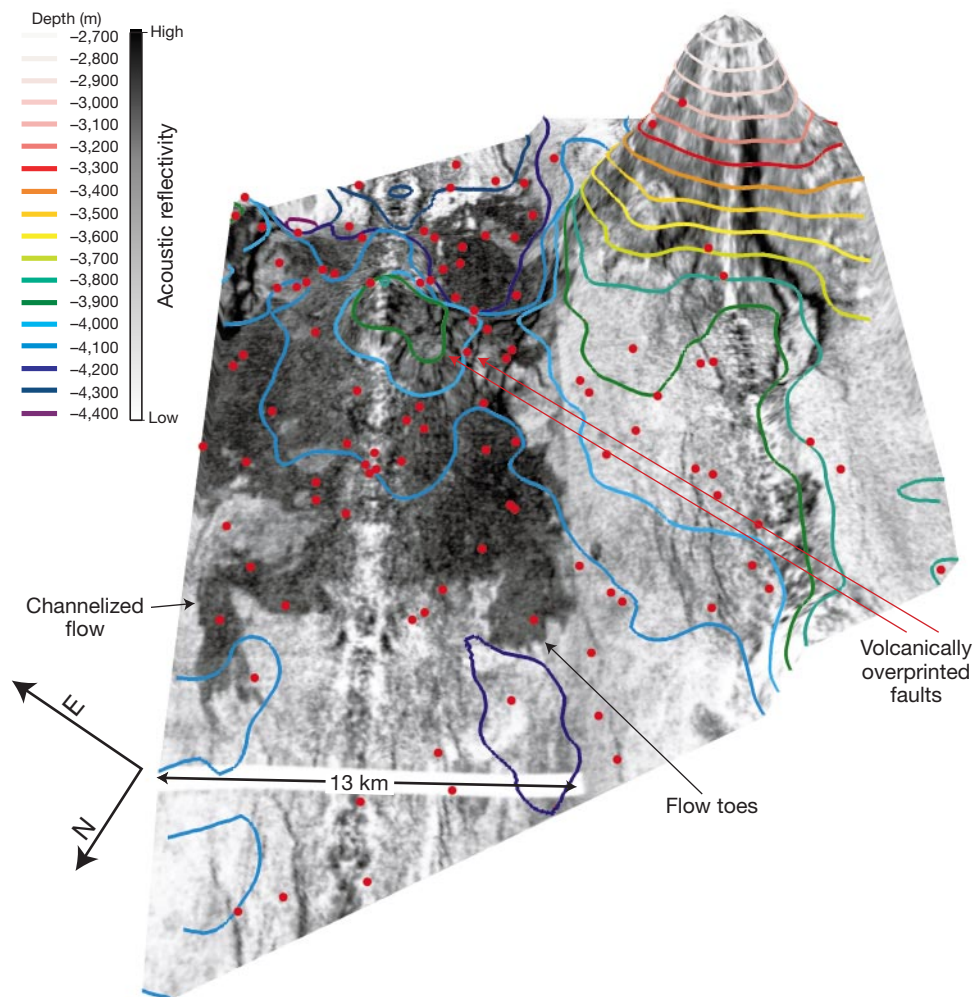


Figure 3 Three-dimensional view of the western volcano looking from west to east. Data presentation is analogous to that described for Fig. 2. The dark, reflective terrain is centred about a close-contoured high having a maximum vertical relief of 500 m. As in Fig. 2, lava channels spill downslope from the volcano, ponding against fault scarps or terminating in flow toes that are characteristic of eruptive processes. Lava on the western volcano is significantly less faulted than lava on the eastern volcano (Fig. 2). The few faults evident on the southern flank of the western volcano are located on a small saddle

between the western volcano and the prominent volcano to its south (right). On the saddle, these faults cut through both strongly and poorly reflective terrain indicating that at least some of the highly reflective terrain has undergone substantial tectonic modification. However, the faults abruptly terminate to west and east of the saddle, suggesting that they have been volcanically overprinted. Red circles show the locations of epicentres for the Gakkel ridge earthquake swarm that ran from January until September in 1999. See Fig. 1 for the location of the western volcano.

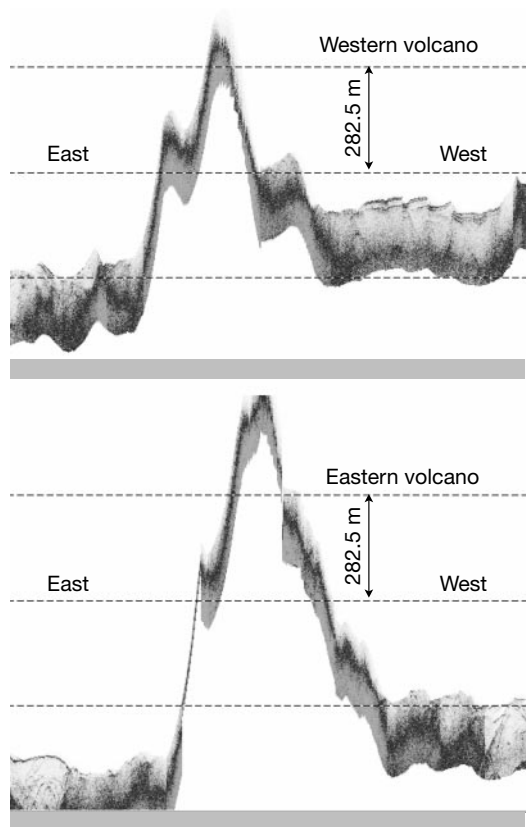


Figure 4 SCAMP sub-bottom data for the western and eastern volcanoes show no evidence of sediment layering on top of the constructs, although layers are apparent adjacent to the flanks of both highs. In these plots the *x*-axis represents time; the USS *Hawkbill* was heading from east to west when the data were collected. Vertical relief is indicated.

provide a revised model for volcanism at the ultraslow-spreading Gakkel ridge, in which voluminous, sustained eruptions focused at discrete sites may have occurred more frequently than previously thought. □

Received 7 June; accepted 13 November 2000.

1. Chen, Y. J. Oceanic crustal thickness versus spreading rate. *Geophys. Res. Lett.* **19**, 753–756 (1992).
2. Bown, J. W. & White, R. S. Variation with spreading rate of oceanic crustal thickness and geochemistry. *Earth Planet. Sci. Lett.* **121**, 435–449 (1994).
3. Sparks, D. W., Parmentier, E. M. & Phipps Morgan, J. Three-dimensional mantle convection beneath a segmented spreading center: Implications for along-axis variations in crustal thickness and gravity. *J. Geophys. Res.* **98**, 21977–21995 (1993).

4. Reid, I. & Jackson, H. R. Oceanic spreading rate and crustal thickness. *Mar. Geophys. Res.* **5**, 165–172 (1981).
5. Miller, C. & Jokat, W. Seismic evidence for volcanic activity discovered in central Arctic. *Eos* **81**, 265–269 (2000).
6. Laxon, S. & McAdoo, D. Arctic Ocean gravity field derived from ERS-1 satellite altimetry. *Science* **265**, 621–624 (1994).
7. Smith, W. H. F. & Sandwell, D. T. Global sea floor topography from satellite altimetry and ship depth soundings. *Science* **277**, 1956–1962 (1997).
8. Chayes, D. *et al.* Swath Mapping the Arctic Ocean from US Navy submarines: installation and performance analysis of SCAMP operations during SCICEX 1998. *Eos* **79**, F854 (1998).
9. DeMets, C., Gordon, R. G., Argus, D. F. & Stein, S. Current plate motions. *Geophys. J. Int.* **101**, 425–478 (1990).
10. Vogt, P. R., Taylor, P. T., Kovacs, L. C. & Johnson, G. L. Detailed aeromagnetic investigation of the Arctic Basin. *J. Geophys. Res.* **84**, 1071–1089 (1979).
11. Coles, R. L. & Taylor, P. T. in *The Arctic Ocean Region, The Geology of North America* Vol. L, 119–132 (eds Grantz, A., Johnson, G. L. & Sweeney, J. F.) (Geological Society of America, Boulder, CO, 1990).
12. Kristoffersen, Y. The Nansen Ridge, Arctic Ocean: Some geophysical observations of the rift valley at slow spreading rate. *Tectonophysics* **89**, 161–172 (1982).
13. Ballard, R. D. & Van Andel, T. H. Morphology and tectonics of the inner rift valley at lat 36° 50' N on the Mid-Atlantic Ridge. *Geol. Soc. Am. Bull.* **88**, 507–530 (1977).
14. Smith, D. K. & Cann, J. R. Constructing the upper crust of the Mid-Atlantic Ridge: A reinterpretation based on the Puna Ridge, Kilauea Volcano. *J. Geophys. Res.* **104**, 25379–25399 (1999).
15. Coakley, B. J. & Cochran, J. R. Gravity evidence of very thin crust at the Gakkel Ridge (Arctic Ocean). *Earth Planet. Sci. Lett.* **162**, 81–95 (1998).
16. Urlick, R. J. *Principles of Underwater Sound* 136–143 (McGraw-Hill, New York, 1983).
17. Thiede, J., Clark, D. L. & Herman, Y. in *The Arctic Ocean Region: The Geology of North America* Vol. L, 427–458 (eds Grantz, A., Johnson, G. L. & Sweeney, J. F.) (Geological Society of America, Boulder, CO, 1990).
18. Macdonald, K. C., Haymon, R. & Shor, A. A 220 km² recently erupted lava field on the East Pacific Rise near lat 8° S. *Geology* **17**, 212–216 (1989).
19. Johnson, H. P. & Helferty, M. The geological interpretation of side-scan sonar. *Rev. Geophys.* **28**, 357–380 (1990).
20. Chadwick, W. W., Embley, R. W. & Fox, C. G. Evidence for volcanic eruption on the southern Juan de Fuca ridge between 1981 and 1987. *Nature* **350**, 416–418 (1991).
21. Hall, L. S. & Sinton, J. M. Geochemical diversity of the large lava field on the flank of the East Pacific Rise at 8° 17' S. *Earth Planet. Sci. Lett.* **142**, 241–251 (1996).
22. McAllister, E. & Cann, J. R. in *Tectonic, Hydrothermal and Geological Segmentation of Mid-Ocean Ridges* (eds MacLeod, C. J., Tyler, P. A. & Walker, C. L.) 29–48 (Geol. Soc. Spec. Pub. 188, Geological Society of America, Boulder, CO, 1996).
23. Stein, R. S., Briole, P., Ruegg, J. C., Tapponnier, P. & Gasse, F. Contemporary, Holocene and Quaternary deformation of the Asal Rift, Djibouti: Implications for the mechanics of slow spreading ridges. *J. Geophys. Res.* **96**, 21789–21806 (1991).
24. Simkin, T. & Howard, K. A. Caldera collapse in the Galapagos Islands, 1968. *Science* **169**, 429–437 (1970).
25. Crane, K. *et al.* Volcanic and seismic swarm events on the Reykjanes Ridge and their similarities to events on Iceland: Results of a rapid response mission. *Mar. Geophys. Res.* **19**, 319–337 (1997).
26. Tolstoy, M., Bohnenstiehl, D. R., Edwards, M. H. & Kurras, G. J. Eruption processes at the ultra-slow spreading Gakkel Ridge. *Eos* **81**, F1346 (2000).

Acknowledgements

We thank the captain (R. Perry), the officers and crew of the USS *Hawkbill* and the scientists and engineers who sailed during SCICEX-98 and SCICEX-99. We also thank D. Chayes, R. Anderson, D. Fornari and K. Macdonald for comments. R. Davis and B. Appelgate provided software and data assistance. P. Johnson created the three-dimensional renderings. This work is the result of M. Langseth's vision. SCAMP was funded by the National Science Foundation, Lamont-Doherty Earth Observatory of Columbia University, the Palisades Geophysical Institution and the governments of Canada, Norway and Sweden. This research was supported by the National Science Foundation.

Correspondence and requests for materials should be addressed to M.E. (e-mail: margo@soest.hawaii.edu).



nature

the human genome

Everyone's genome

"The human genome underlies the fundamental unity of all members of the human family, as well as the recognition of their inherent dignity and diversity. In a symbolic sense, it is the heritage of humanity."
Universal Declaration on the Human Genome and Human Rights (http://www.unesco.org/human_rights/hrbc.htm)

Humans are much more than simply the product of a genome, but in a sense we are, both collectively and individually, defined within the genome. The mapping, sequencing and analysis of the human genome is therefore a fundamental advance in self-knowledge; it will strike a personal chord with many people. And application of this knowledge will, in time, materially benefit almost everyone in the world.

It is with particular pleasure, therefore, that we present this special section of *Nature* on the human genome. It provides a comprehensive overview of current knowledge, in three sections:

- The **News and Views** articles provide the context, communicate the excitement, and present a critical evaluation of the papers.
- The **Analysis** section delves into the opportunities offered to biologists by the genome and explores the approaches and tools needed to exploit these opportunities.
- The **Research** section comprises a set of papers that, for the most part, report the findings of the international consortia of scientists that make up the Human Genome Project. More than 2,500 authors from 20 laboratories contributed to these papers.

Recognizing the need for engaging **educational material**, the issue is accompanied by a wall chart, "The Geography of the Genome", and an educational CD-ROM produced by the US National Human Genome Research Institute (and co-sponsored by *Nature*). It includes a historical timeline of genetics, the background to the Human Genome Project, and teaching aids to explain the fundamental principles of genetics.

A suite of interwoven maps is one of the features of the issue. The key map for the entire project is the whole-genome clone-based physical map (p. 934) constructed by the International Human Genome Mapping Consortium. It provided the scaffold upon which the sequence was assembled. The cytogenetic map, described on p. 953, plots landmarks across the genome and anchors the physical map to the underlying chromosomal positions. A comparison of the genetic and physical maps (p. 951) charts the rate of recombination — or exchange between each pair of our 46 chromosomes — that occurs as the genome is passed on through generations. Finally, the International SNP [single nucleotide polymorphism] Map Working Group documents 1.42 million polymorphic sites in the genome (p. 928), providing for the first time a variant in virtually every gene and in each genomic region. This map of human variation will facilitate efforts to reconstruct our evolutionary history and dissect the genetic basis of human traits and disease.

Analysis of the draft genome sequence itself (p. 860) provides the first panoramic view of the landscape of our genome. We learn the extent to which it has been colonized by parasitic DNA elements. In times past, these elements underwent massive proliferation, playing an important role in shaping the evolution of the human genome. Another notable feature is the much lower gene tally than anticipated, which indicates that human complexity does not arise solely from the number of genes.

Although computational predictions can provide 'best guesses' on gene number and structure, definitive answers require experimental data on gene expression, in terms of both temporal and positional analyses of gene products. As a step towards this, microarray technology has been used to validate gene predictions and more accurately define gene structures (p. 922).

One principle at the heart of the Human Genome Project, reflected in the Universal Declaration on the Human Genome and Human Rights, is free and unlimited access to the sequence. In keeping with this principle, the entire content of this section, plus additional features and commentary, is available without restriction (<http://www.nature.com/genomics>).

We welcome feedback on any aspect of this publication by e-mail at genome@nature.com.

Carina Dennis Senior Editor
Richard Gallagher Chief Biology Editor
Philip Campbell Editor-in-chief

Our genome unveiled

David Baltimore

The draft sequences of the human genome are remarkable achievements. They provide an outline of the information needed to create a human being and show, for the first time, the overall organization of a vertebrate's DNA.

I've seen a lot of exciting biology emerge over the past 40 years. But chills still ran down my spine when I first read the paper that describes the outline of our genome and now appears on page 860 of this issue¹. Not that many questions are definitively answered — for conceptual impact, it does not hold a candle to Watson and Crick's 1953 paper² describing the structure of DNA. Nonetheless, it is a seminal paper, launching the era of post-genomic science.

This milestone of biology's megaproject is the long-promised draft DNA sequence from the International Human Genome Sequencing Consortium (the public project). The sequence itself is available to all those connected to the Internet³. In the paper in this issue, we are presented with a description of the strategy used to decipher the structures of the huge DNA molecules that constitute the genome, and with analyses of the content encoded in the genome. It is the achievement of a coordinated effort involving 20 laboratories and hundreds of people around the world. It reflects the scientific community at its best: working collaboratively, pooling its resources and skills, keeping its focus on the goal, and making its results available to all as they were acquired.

Simultaneously, another draft sequence is being published⁴. It is less freely available because it was generated by a company, Celera Genomics, that hopes to sell the information. This week's *Science* contains an account of the history of that project and the analyses of its data, while another of the papers in this issue contains a comparison of the quality of the two sequences⁵. To those who saw this as a competitive sport, the papers make it appear to be roughly a tie. However, it is important to remember that Celera had the advantage of all of the public project's data. Nevertheless, Celera's achievement of producing a draft sequence in only a year of data-gathering is a testament to what can be realized today with the new capillary sequencers, sufficient computing power and the faith of investors.

Answers

What have we learned from all of these AGCTs? The best way to answer the question is to read the analytical sections of the papers. I will only make some general comments. It is important to remember that no statements can be made with high precision because the draft sequences have holes and imperfections, and the tools for analysis remain limited (as described in a further paper⁶ in this issue, page 828). However, the answers provided by the draft will be of interest to many investigators, and the value of having the draft published in its imperfect form is unquestionable.

The sequences are about 90% complete for the euchromatic (weakly staining, gene-rich) regions of the human chromosomes. The estimated total size of the genome is 3.2 Gb (that is gigabases, the latest escalation of units needed to contain the fruits of modern technology). Of that, about 2.95 Gb is euchromatic. Only 1.1% to 1.4% is sequence that

actually encodes protein; that is just 5% of the 28% of the sequence that is transcribed into RNA. Over half of the DNA consists of repeated sequences of various types: 45% in four classes of parasitic DNA elements, 3% in repeats of just a few bases, and about 5% in recent duplications of large segments of DNA. The amounts in the first and third classes will certainly grow as our ability to characterize them increases in effectiveness and we examine the darkly staining, heterochromatin regions of chromosomes. As the co-discoverer of reverse transcriptase (the enzyme that reverses the common mode of information transfer from DNA to RNA), I find it striking that most of the parasitic DNA came about by reverse transcription from RNA. In places, the genome looks like a sea of reverse-transcribed DNA with a small admixture of genes.

Repeats

By contrast, the puffer fish — another vertebrate — has a genome that contains very few repeats. But it encodes a perfectly functional creature, so it seems likely that most of the repeats are simply parasitic, selfish DNA elements that use the genome as a convenient host. People call this 'junk DNA', but from the DNA's point of view it deserves more respect. In most places in the human genome the selfish elements are tolerated, and in some places — near the ends of chromosomes, for instance, or near the chromosome constrictions called centromeres — it builds up to form huge segments. However, the repeated DNA may have both negative and positive effects. For instance, the paucity of repeats in certain highly regulated regions of the genome suggests that insertions there can disrupt gene regulation and are deleterious. Conversely, the enrichment of the so-called Alu class of repeated sequences in the gene-rich, high-GC regions of the genome implies that they have a positive function. The repeats can also be fodder for evolving new functions and act as loci for gene rearrangements.

In humans, virtually all of the parasitic DNA repeats seem old and enfeebled, with little evidence of continuing reinversions. However, there has been very little evolutionary scouring of these repeats from the human genome, making it a rich record of evolutionary history. The mouse genome, by contrast, has many actively reinserting parasitic sequences and is scoured more intensely, making it a much younger and more dynamic genome. This difference might reflect the shorter generation time of mice or something about their physiology, but I find it an intriguingly enigmatic observation.

Much of what we learn about the global organization of the genome is an elaboration of previous notions. For instance, we knew that the genome had regions with a relatively high content of GC bases and regions high in AT, but now we have a very complete appreciation of this architecture. What maintains the patchiness of the GC/AT ratio in the genome remains an unanswered question. As was expected,

most genes are located outside the heterochromatic regions; interestingly, however, in regions of the genome rich in GC bases, the gene density is greater and the average intron size is lower. These introns — made up of largely meaningless sequence that breaks up the protein-coding sequences (exons) of genes — are much longer in human DNA than in the genomes previously sequenced. Their dilution of the coding sequence is one element that makes finding genes by computer so difficult in human DNA.

A major interest of the genome sequence to many biologists will be the opportunity it provides to discover new genes in their favourite systems — for instance, cell biologists will search for new genes for signalling proteins, and neurobiologists will look for

new ion channels. This data-mining exercise was carried out by various groups which report their initial findings in papers that appear on pages 824–859 of this issue. They found some new and interesting genes, but surprisingly few, and occasionally could not find the full extent of genes that they knew were there. The paucity of discoveries reflects their concentration on systems that were previously heavily studied.

Gene-regulatory sequences are now there for all to see, but initial attempts to find them were also disappointing. This is where the genomic sequences of other species — in which the regulatory sequences, but not the functionally insignificant DNA, are likely to be much the same — will open up a cornucopia. Basically, the human sequence at its

present level of analysis allows us to answer many global questions fairly well, but the detailed questions remain open for the future.

What interested me most about the genome? The number of genes is high on the list. The public project estimates that there are 31,000 protein-encoding genes in the human genome, of which they can now provide a list of 22,000. Celera finds about 26,000. There are also about 740 identified genes that make the non-protein-coding RNAs involved in various cell housekeeping duties, with many more to be found. The number of coding genes in the human sequence compares with 6,000 for a yeast cell, 13,000 for a fly, 18,000 for a worm and 26,000 for a plant. None of the numbers for the multicellular organisms is highly

Genome speak

Allele Humans carry two sets of chromosomes, one from each parent.

Equivalent genes in the two sets might be different, for example because of *single nucleotide polymorphisms*. An allele is one of the two (or more) forms of a particular gene.

Bacterial artificial chromosome (BAC) A chromosome-like structure, constructed by genetic engineering, that carries genomic DNA to be *cloned*.

Centromere Chromosomes contain a compact region known as a centromere, where sister chromatids (the two exact copies of each chromosome that are formed after replication) are joined.

Cloning The process of generating sufficient copies of a particular piece of DNA to allow it to be sequenced or studied in some other way.

Complementary DNA (cDNA) A DNA sequence made from a *messenger RNA* molecule, using an enzyme called reverse transcriptase. cDNAs can be used experimentally to determine the sequence of messenger RNAs after their introns (non-protein-coding sections) have been *spliced* out.

Conservation Genes that are present in two distinct organisms are said to be conserved. Conservation can be detected by measuring the similarity of the two sequences at the base (RNA or DNA) or amino-acid (protein) level. The more similarities there are, the more highly conserved the two sequences.

Euchromatin The gene-rich regions of a genome (see also *heterochromatin*).

Eukaryote An organism whose cells have a complex internal structure, including a nucleus. Animals, plants and fungi are all eukaryotes.

Expressed sequence tag (EST) A short piece of DNA sequence corresponding to a fragment of a *complementary DNA* (made from a cell's *messenger RNA*). ESTs have been used to hunt for genes, so hundreds of thousands are present in sequence databases.

Genome The complete DNA sequence of an organism.

Genotype The set of genes that an individual carries; usually refers to the particular pair of alleles (alternative forms of a gene) that a person has at a given region of the genome.

Haplotype A particular combination of alleles (alternative forms of genes) or sequence variations that are closely linked — that is, are likely to be inherited together — on the same chromosome.

Heterochromatin Compact, gene-poor regions of a genome, which are enriched in simple sequence repeats. As it can be impossible to *clone*, heterochromatin is often ignored when calculating the percentage of a genome that has been sequenced. Heterochromatin was originally identified as regions of the genome that stained differently to euchromatin (gene-rich regions).

Introns and exons Genes are *transcribed* as continuous sequences, but only some segments of the resulting *messenger RNA* molecules contain information that codes for the gene's protein product. These segments are called exons. The regions between exons are known as introns, and are *spliced* from the RNA before the product is made.

Long and short arms The regions either side of the centromere, a compact part

of a chromosome, are known as arms. As the centromere is not in the centre of the chromosome, one arm is longer than the other.

Messenger RNA (mRNA) Proteins are not synthesized directly from genomic DNA. Instead, an RNA template (a precursor mRNA) is constructed from the sequence of the gene. This RNA is then processed in various ways, including *splicing*. Spliced RNAs destined to become templates for protein synthesis are known as mRNAs.

Mutation An alteration in a genome compared to some reference state. Mutations do not always have harmful effects.

Phenotype The observable properties and physical characteristics of an organism.

Polymorphism A region of the genome that varies between individual members of a population. To be called a polymorphism, a variant should be present in a significant number of people in the population.

Prokaryote A single-celled organism with a simple internal structure and no nucleus. Bacteria and archaeobacteria are prokaryotes.

Proteome The complete set of proteins encoded by the *genome*.

Pseudogene A region of DNA that shows extensive similarity to a known gene, but which cannot itself function, either because it has lost the signal required for *transcription* (the promoter sequence) or because it carries mutations that prevent it from being *translated* into protein.

Recombination The process by which DNA is exchanged between pairs of equivalent chromosomes during egg and sperm formation. Recombination has the effect of making the chromosomes of the offspring distinct from those of the parents.

Restriction endonuclease An enzyme that cleaves DNA at every location at which a particular short sequence occurs. Different types of restriction endonuclease cleave at different target sequences.

Single nucleotide polymorphism (SNP) A *polymorphism* caused by the change of a single nucleotide. Most genetic variation between individual humans is believed to be due to SNPs.

Splicing The process that removes introns (non-protein-coding portions) from *transcribed* RNAs. Exons (protein-coding portions) can also be removed. Depending on which exons are removed, different proteins can be made from the same initial RNA or gene. Different proteins created in this way are 'splice variants' or 'alternatively spliced'.

Transcription The process of copying a gene into RNA. This is the first step in turning a gene into a protein, although not all transcripts lead to proteins.

Transcriptome The complete set of RNAs *transcribed* from a genome.

Translation The process of using a *messenger RNA* sequence to build a protein. The messenger RNA serves as a template on which transfer RNA molecules, carrying amino acids, are lined up. The amino acids are then linked together to form a protein chain.

Peer Bork and Richard Copley

accurate because of the limitations of gene-finding programs. But unless the human genome contains a lot of genes that are opaque to our computers, it is clear that we do not gain our undoubted complexity over worms and plants by using many more genes. Understanding what does give us our complexity — our enormous behavioural repertoire, ability to produce conscious action, remarkable physical coordination (shared with other vertebrates), precisely tuned alterations in response to external variations of the environment, learning, memory... need I go on? — remains a challenge for the future.

Complexity

Where do our genes come from? Mostly from the distant evolutionary past. In fact, only 94 of 1,278 protein families in our genome appear to be specific to vertebrates. The most elementary of cellular functions — basic metabolism, transcription of DNA into RNA, translation of RNA into protein, DNA replication and the like — evolved just once and have stayed pretty well fixed since the evolution of single-celled yeast and bacteria. The biggest difference between humans and worms or flies is the complexity of our proteins: more domains (modules) per protein and novel combinations of domains. The history is one of new architectures being built from old pieces. A few of our genes seem to have come directly from bacteria, rather than by evolution from bacteria — apparently bacterial genomes can be direct donors of genes to vertebrates. So DNA chimaeras consisting of the genes from several organisms can arise naturally as well as artificially (opponents of 'genetically modified foods' take note).

The most exciting new vista to come from the human genome is not tackling the question "What makes us human?", but addressing a different one: "What differentiates one organism from another?". The first question, imprecise as it is, cannot be answered by staring at a genome. The second, however, can be answered this way because our differences from plants, worms and flies are mainly a consequence of our genetic endowments. The Celera team⁴ presents the more detailed analysis of the numbers of different protein motifs and protein types, in extensive tables. From them, it is easy to see what types of proteins and motifs have been amplified for specific types of organisms. In vertebrates, not surprisingly, we see elaboration and the *de novo* appearance of two types of genes: those for specific vertebrate abilities (such as neuronal complexity, blood-clotting and the acquired immune response), and those that provide increased general capabilities (such as genes for intra- and intercellular signalling, development, programmed cell death, and control of gene transcription). Someday soon we will have the mouse genome, and then those of fish and dogs, and probably the kangaroo genome from the

Australians. Each of these will fill in a piece of the evolutionary puzzle and will provide exciting comparisons.

We wait with bated breath to see the chimpanzee genome. But knowing now how few genes humans have, I wonder if we will learn much about the origins of speech, the elaboration of the frontal lobes and the opposable thumb, the advent of upright posture, or the sources of abstract reasoning ability, from a simple genomic comparison of human and chimp. It seems likely that these features and abilities have mainly come from subtle changes — for example, in gene regulation, in the efficiency with which introns are spliced out of RNA, and in protein-protein interactions — that are not now easily visible to our computers and will require much more experimental study to tease out. Another half-century of work by armies of biologists may be needed before this key step of evolution is fully elucidated.

What is next? Lots of hard work, but with new tools and new aims. First, we have to stay the course and get the most precise representation of the genome that we can: this is a matter of filling the cracks, cleaning up the errors, and getting rid of the uncertainties that plague each of the analytical methods. Second, we need to see more genomes, with each one giving us a deeper insight into our own. Third, we need to learn how to take advantage of this book of life. Tools for scanning the activity levels of genes in different cells, tissues and settings are becoming available and are already revolutionizing how we do biological investigation. But we will have to move back from the general to the particular, because each gene is a story in itself and its full significance can be learned only from concentrating on its particular properties.

Fourth, we need to turn our new genomic information into an engine of pharmaceuti-

cal discovery. Individual humans differ from one another by about one base pair per thousand. These 'single nucleotide polymorphisms' (SNPs) are markers that can allow epidemiologists to uncover the genetic basis of many diseases. They can also provide information about our personal responses to medicines — in this way, the pharmaceutical industry will get new targets and new tools to sharpen drug specificity. Moreover, the analysis of SNPs will provide us with the power to uncover the genetic basis of our individual capabilities such as mathematical ability, memory, physical coordination, and even, perhaps, creativity.

Biology today enters a new era, mainly with a new methodology for answering old questions. Those questions are some of the deepest and simplest: "Daddy, where did I come from?"; "Mommy, why am I different from Sally?". As these and other questions get robust answers, biology will become an engine of transformation of our society. Instead of guessing about how we differ one from another, we will understand and be able to tailor our life experiences to our inheritance. We will also be able, to some extent, to control that inheritance. We are creating a world in which it will be imperative for each individual person to have sufficient scientific literacy to understand the new riches of knowledge, so that we can apply them wisely.

David Baltimore is at the California Institute of Technology, 1200 East California Boulevard, Mail Code 204-31, Pasadena, California 91125, USA.
e-mail: baltimo@caltech.edu

1. International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
2. Watson, J. D. & Crick, F. H. C. *Nature* **171**, 737–738 (1953).
3. <http://genome.cse.ucsc.edu/>
4. Venter, J. C. *et al.* *Science* **291**, 1304–1351 (2001).
5. Aach, J. *et al.* *Nature* **409**, 856–859 (2001).
6. Birney, E., Bateman, A., Clamp, M. E. & Hubbard, T. J. *Nature* **409**, 827–828 (2001).

The maps

Clone by clone by clone

Maynard V. Olson

The public project's sequencing strategy involved producing a map of the human genome, and then pinning sequence to it. This helps to avoid errors in the sequence, especially in repetitive regions.

This issue of *Nature* celebrates a halfway point in the implementation of the 'map first, sequence later' strategy adopted by the Human Genome Project in the mid-1980s¹. The results suggest that the strategy was basically sound. It led, as hoped, to a project that could be distributed internationally across many genome-sequencing centres, and that would allow sequenced fragments of the human genome to be anchored to mapped genomic landmarks long before the complete sequence coalesced

into one long string of Gs, As, Ts and Cs.

The centrepiece of the suite of mapping papers in this issue is on page 934, where the International Human Genome Mapping Consortium describes a 'clone-based' physical map of the human genome². A map like this not only charts the genome, giving a structure on which to hang sequence data, but also provides a starting point for sequencing. Figure 1 shows the basics of the approach. I drew this figure in 1981, using India ink and a Leroy lettering set. Both

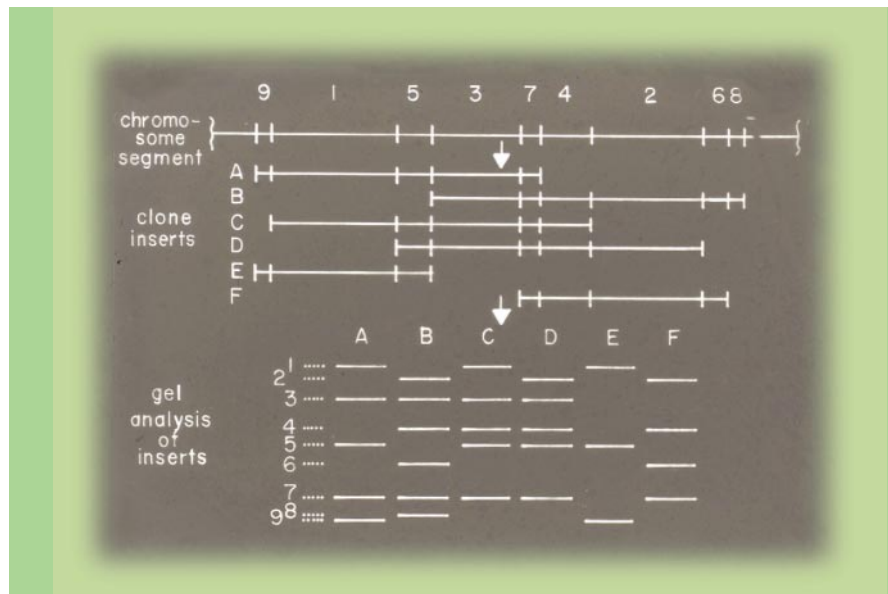


Figure 1 Clone-based physical mapping. The top line shows the location of 'restriction' sites (vertical bars) in a particular region of the genome. Restriction sites are places at which a site-specific restriction endonuclease cleaves DNA. The fragments produced by cleavage at every possible point in this region are numbered 1 to 9. Below the line are several clones with random end points, labelled A to F. Clones are produced by first partially digesting many copies of the genome with different restriction endonucleases; the resulting large segments are then inserted into bacteria and replicated (cloned). Each clone is digested with a restriction endonuclease, and the resulting fragments are separated, by size, on an electrophoretic gel ('gel analysis of inserts'). This process yields a distinctive pattern ('fingerprint') for each clone. The map-assembly problem requires working backwards (upwards in this figure) from the fingerprints to a clone-overlap map and restriction-site map of the chromosome segment. To finish the analysis of this region of the genome, the natural choice of clones to sequence would be A and B.

graphical and mapping technologies have come a long way since then, but the principles behind clone-based physical mapping have not changed.

The clone-based approach works as follows. Many copies of the genome are cut up into segments of about 150,000 base pairs by partial digestion with site-specific restriction endonucleases — enzymes that cleave DNA in specific places. ('Partial' digestion means that the reaction is not carried out for long enough to allow every possible cleavage to be made.) The large DNA segments are plugged into bacterial 'artificial chromosomes' (BACs) and inserted into bacteria, where they are copied exactly each time the bacteria divide. The process produces 'clones' of identical DNA molecules that can be purified for further analysis. Next, each clone is completely digested with a restriction endonuclease, chosen to produce a characteristic pattern of small fragments, or a 'fingerprint', for each clone. Comparison of the patterns reveals overlap between the clones, allowing them to be lined up in order, while the sites in the genome at which the restriction endonuclease cleaves are charted. The result is a physical map.

Individual BAC clones are then sheared into smaller fragments and cloned; the resulting 'small-insert' subclones are sequenced. The sequence of an individual

BAC clone is assembled from the sequences of an 'oversampled' set of subclones (in other words, enough subclones are sequenced to ensure that each part of the original clone is analysed several times). Finally, the whole genome sequence is assembled by melding together the sequences of a set of BACs that spans the genome.

This approach is similar to that used in the 1980s and early 1990s to map and sequence the genomes of the nematode *Caenorhabditis elegans* and the yeast *Saccharomyces cerevisiae*^{3,4}. What is new in the human project is its staggering scale, and the speed with which it has been completed. By way of comparison, although the nematode and yeast genomes are, respectively, only 3% and 0.5% the size of the human genome, these early mapping projects spanned the better part of a decade, as opposed to two years for the much larger human project.

One weakness of clone-based physical mapping is that the maps often have poor continuity. For example, there is not always a BAC clone to cover every part of the genome; and overlaps between clones can be obscured by data errors or the presence of large-scale repeats in the genome. The current map² has more than 1,000 discontinuities. These will cause some difficulties as the Human Genome Project moves to its next phase, which will involve ensuring accuracy and

filling in any gaps in the sequence. Nonetheless, the current map typically maintains continuity for several million base pairs at a stretch. These continuous segments are big enough to allow the clone-based map to be overlaid on various lower-resolution maps. In this way, the mapped segments can be ordered and orientated, much as a discontinuous patchwork of high-resolution maps of the Earth's surface can be orientated by overlaying them on a satellite photograph of the whole Earth.

Two particularly interesting low-resolution maps are the genetic and cytogenetic maps, on pages 951 and 953 of this issue^{5,6}. The genetic map⁵ is based on the probability of the occurrence of recombination — the swapping of corresponding, nearly identical segments of DNA between maternally and paternally derived chromosomes as the genome is passed from one generation to the next. The cytogenetic map⁶ is based on subtle variations in the staining properties of different regions of the genome, as viewed by light microscopy. Yet more papers describe different approaches to clone-based mapping⁷⁻⁹. These methods were applied to particular chromosomes simply because different sequencing centres chose to rely on the whole-genome map to different degrees.

But was all this cartography even necessary? Another draft of the human genome sequence is described in this week's *Science* by Celera Genomics¹⁰. This group adopted a different approach, which involved preparing small-insert clones directly from genomic DNA rather than from mapped BACs. The major rationale for the BAC-by-BAC approach² was to make easier the finishing phase of the Human Genome Project, which lies ahead. The consortium now plans to upgrade the 30,000 BAC sequences by sequencing more subclones from each BAC (the 'topping-up' phase) and then resolving internal gaps and discrepancies (the 'finishing' phase). Segmenting the finishing phase into BAC-sized portions provides an enormous advantage in dealing with blocks of sequence that are repeated at many different places within the genome. The power of this strategy is nicely illustrated by the mapping of the Y chromosome, whose repetitive structure is unusually complex (page 943 of this issue¹¹).

Nature readers should not expect any real answer to the question of which of these two approaches is the better one. But it is likely that the only players still on the field when the toughest finishing issues are confronted will be the public consortium's BAC brigade. In the future, as genome sequencing moves on to other mammals, the context will have changed; the human sequence will provide an invaluable guide to assembling long stretches of sequence that are shared among all mammalian genomes. So the sequencing of the human genome is likely to be the only large

sequencing project carried to completion by the methods described in this issue. Genome sequencing will get easier from here.

Looking ahead, there are two threats to producing a quality finished product. One is simple exhaustion on the part of the consortium's members: each new round of press conferences announcing that the human genome has been sequenced saps the morale of those who must come to work each day actually to do what they read in the newspapers has already been done.

We may also expect to hear the argument that the current sequence is good enough for most purposes, and that remaining problems should be resolved by users as the need for accurate sequence in specific regions arises. What we have now is certainly a lot better than what we had yesterday. But biologists in the future will be comparing vast data sets to the reference sequence of the human genome. They must be able to do so with confidence that the discrepancies they encounter are due to the limitations of their

own data or, more interestingly, to biology. They should not need to expend time, energy and imagination compensating for a failure now to pursue the Human Genome Project to a grand conclusion. We must move on and finish the job, even as the bright lights of media attention shift elsewhere. ■

Maynard V. Olson is in the Departments of Medicine and Genetics, Fluke Hall, Mason Road, University of Washington, Seattle, Washington 98195-2145, USA.

e-mail: mvo@u.washington.edu

1. National Research Council *Mapping and Sequencing the Human Genome* (National Academy Press, Washington DC, 1988).
2. The International Human Genome Mapping Consortium *Nature* **409**, 934–941 (2001).
3. Coulson, A., Sulston, J., Brenner, S. & Karn, J. *Proc. Natl Acad. Sci. USA* **83**, 7821–7825 (1986).
4. Olson, M. V. *et al. Proc. Natl Acad. Sci. USA* **83**, 7826–7830 (1986).
5. Yu, A. *et al. Nature* **409**, 951–953 (2001).
6. The BAC Resource Consortium *Nature* **409**, 953–958 (2001).
7. Montgomery, K. T. *et al. Nature* **409**, 945–946 (2001).
8. Brůls, T. *et al. Nature* **409**, 947–948 (2001).
9. Bentley, D. R. *et al. Nature* **409**, 942–943 (2001).
10. Venter, J. C. *et al. Science* **291**, 1304–1351 (2001).
11. Tilford, C. A. *et al. Nature* **409**, 943–945 (2001).

For the other draft, that produced by Celera Genomics², a variety of methods suggest that between 88% and 93% of the euchromatin has been sequenced and assembled. But direct comparison of these numbers with the public consortium's draft is almost impossible — different procedures and measures were used to process the data and to estimate accuracy. Both projects also have sequence data that were not used in the assembly process, raising the real level of coverage by a few percentage points.

These numbers might seem rather arbitrary, but even when the first genome of an animal species was published⁵, it was clear that simple, practical finish lines do not exist (Box 1, Fig. 1). The present level of coverage of the human genome reflects the point where a shift of focus occurs, from sequencing the genome many times over to producing a high-quality, continuous sequence⁶. There is some way to go yet.

Essentially, 'rough draft' refers to the fact that the sequences are not continuous — there are gaps (Box 1). If there are too many gaps, it can be impossible to order and orientate the many small strings of bases that are the raw products of genome sequencing. This might, for example, hamper projects that seek to identify genes involved in inherited diseases. A first step to finding such genes is to work out which region of which chromosome they are on. The complete genome sequence should be immensely useful for the next step — identifying the relevant gene at that region. But gaps and errors in ordering and placing the strings of sequence will make this difficult.

Another problem of incompleteness is that it is difficult to make definitive

The draft sequences

Filling in the gaps

Peer Bork and Richard Copley

Two rough drafts of the human genome sequence are now published. Completion of the sequences lies ahead, but the implications for studying human diseases and for biotechnology are already profound.

With the publication of the human genome sequence — described and analysed on page 860 of this issue¹ and in this week's *Science*² — we cross a border on the route to a better understanding of our biological selves. But unlike the previously published sequences of human chromosomes 21 and 22 (refs 3,4), the present sequences of the whole human genome are not considered complete. The bulk of the data make up what is called a 'rough draft'. So what is all the fuss about? What exactly does 'rough draft' mean, and what can we learn from sequences such as this?

In the draft from the publicly funded International Human Genome Sequencing Consortium¹, around 90% of the gene-rich — euchromatic — portion of the genome has been sequenced and 'assembled', the term used to describe the process of using a computer to join up bits of sequence into a larger whole. Each base pair of this 90% was sequenced four times on average, ensuring reasonable precision. Only about a quarter of the whole genome is considered 'finished' — another bit of genomics jargon, which basically means that each base pair has been sequenced eight to ten times on average, with gaps in the sequence existing only because of the limitations of present technology. Nonetheless, the sequence of base pairs in

the draft is very accurate, and is unlikely to change much; 91% of the euchromatin sequenced has an error rate of less than one base in 10,000 (ref. 1).

Box 1 What makes a completely sequenced genome?

When is sequencing work on a genome complete? No genome for a eukaryotic organism — roughly, those organisms whose cells contain a nucleus — has been sequenced to 100%. There are regions, often highly repetitive, that are difficult or impossible to clone (one of the initial steps in a sequencing project) or sequence with current technology. Fortunately, such regions are expected to contain relatively few protein-coding genes^{4,10}.

The extent of these regions varies widely in different species. So, rather than applying a universal gold standard, each sequencing project has made pragmatic decisions as to what constitutes a sufficient level of coverage for a particular genome. For example, as much as one-third of the sequence of the fruitfly *Drosophila melanogaster* was not stable in the cloning systems used, and so was not sequenced. But 97% of the so-called euchromatic portion — where most genes are thought to reside — was sequenced¹¹ (Fig. 1).

For the human genome, one definition of 'finished' is that fewer than one base in 10,000 is incorrectly assigned⁶; more than 95% of the euchromatic regions are sequenced; and each gap is smaller than 150 kilobases¹². Such standards represent realistic goals given current technology. By this standard, over a quarter of the public consortium's sequence¹ is considered finished at present, including the previously published long arms of chromosomes 21 and 22 (refs 3,4; Fig. 1). The Celera sequences of chromosomes 21 and 22 are slightly more gappy than those from the public consortium, but the converse seems to be true for the other chromosomes². But again, as different protocols were used, it is not easy to compare the overall status of the two assemblies. In the longer term, as much of the heterochromatin — which is harder to sequence, and contains few genes — as possible must be sequenced, because we might otherwise miss important features.

P.B. & R.C.

Organism	Year	Millions of bases sequenced	Total coverage (%)	Coverage of euchromatin (%)	Predicted number of genes	Number of genes per million bases sequenced
<i>Saccharomyces cerevisiae</i>	1996	12	93	100	5,800	483
<i>Caenorhabditis elegans</i>	1998	97	99	100	19,099	197
<i>Drosophila melanogaster</i>	2000	116	64	97	13,601	117
<i>Arabidopsis thaliana</i>	2000	115	92	100	25,498	221
Human chromosome 21	2000	34	75	100	225	7
Human chromosome 22	1999	34	70	97	545	16
Human genome rough draft (public sequence)	2001	2,693	84	90	31,780	12
Human genome rough draft (Celera sequence)	2001	2,654	83	88–93	39,114	15

Figure 1 Sequenced eukaryotic genomes. Total coverage uses an estimate of the total genome size and includes heterochromatin (condensed genomic areas that were originally characterized by staining techniques, and are thought to be highly repetitive and gene-poor). The gene-rich areas make up euchromatin. Gene numbers are taken from the original sequence publications^{1–6,14,15}; most numbers have since changed slightly and different sources give different estimates depending on protocols. The data for the public consortium's rough draft of the human genome are taken from ref. 1, Table 8, page 872. The estimate of total coverage for the Celera data is based on the public consortium's estimate of the full genome size (3,200 million base pairs); the percentage of euchromatin covered is taken from ref. 2. The predicted numbers of human genes are discussed further in the text.

statements about which genes are unique to other species and do not have relatives in the human genome. So it might be prudent not to place too much emphasis on such 'missing' genes at this stage. Even so, they are running out of places to hide, particularly because the level of coverage of the human genome is probably higher than reported here^{1,2} — there are other chunks of unassembled genome sequence in public databases, such as in independent collections of so-called expressed sequence tags.

But ensuring high quality and high coverage are only two aspects of producing a finished genome. For most biologists, the real interest is in the genes themselves. Here, the picture is less rosy, although the problems are caused not so much by the draft nature of the sequence as by the difficulty in finding genes among the other genomic DNA (Box 2).

Even coming up with a rough count of the number of genes is not straightforward. The public consortium's initial set contains about 32,000 genes, made up of around 15,000 known genes and 17,000 predictions. But these 32,000 genes are estimated to come from around 24,500 actual genes — some predicted genes could be 'pseudogenes', or just fragments of real genes. On the other hand, the sensitivity of prediction tends to be only about 60%, so it is reasonable to assume that another 6,800 or so genes (40% of

17,000) have been overlooked. This is how the present estimate of about 31,000 genes (6,800 plus 24,500) was reached¹. Celera predicts that there are around 39,000 genes, but warns that the evidence for some 12,000 of these is weak². The two groups use different gene-identification techniques, so these numbers are not directly comparable. Minor changes in procedures or data could alter either figure considerably. For example, such changes led to a recent estimate being lowered^{7,8} from 120,000 to fewer than 81,000 — and both now seem untenable. Much is a matter of interpretation.

Fortunately, there is every reason to believe that the quality of gene prediction will rapidly improve, and an experimental technique for doing so is discussed on page 922 (ref. 9). With the sequencing of the genomes of other vertebrates, our ability to detect genes by their similarity to known sequences will get better. This is because, thanks to natural selection, gene sequences tend to be altered less during evolution than the DNA surrounding them. In a couple of years we should have at least a more complete list of testable gene candidates.

Despite all this, the information now available has profound implications. For example, there are already many heavily hunted disease-associated genes that have been identified using the public draft (ref. 1, Table 26, page 912). Together with studies of

Box 2 When is a predicted gene a gene?

How many genes are encoded in the human genome? This is a simple question without — as yet — a straightforward answer¹³. The density of genes in the human genome is much lower than for any other genome sequenced so far (Fig. 1), making it particularly difficult to predict where genes are.

Both Celera and the public sequencing consortium used computational algorithms to model genes and make predictions, but such methods are far from perfect. Not only can the start and end positions of a predicted gene be wrong, but exons (the coding parts of a gene) can be missed entirely or wrongly predicted to exist. To reduce this latter effect, the public sequencing consortium required the exons of predicted genes to be 'confirmed', by showing significant similarity to a known sequence (DNA or protein) in a database. But this requirement might be too conservative, making it difficult to predict the presence of new gene families. Celera has required similar confirmation of predictions, but its mouse-genome sequencing project may have provided evidence for further vertebrate-specific genes.

Spurious prediction is also a problem. All genes are expressed by being copied (transcribed) into messenger RNA; most messenger RNAs are then translated into proteins. But even evidence that a stretch of DNA is transcribed does not definitively show that stretch to be a gene. We do not know how efficiently cells control transcription; indeed, it seems likely that non-gene DNA sequences are transcribed relatively frequently¹². Nor do we know how well the cell identifies transcripts that cannot be translated into a functioning protein. Moreover, proteins that cannot serve any useful function (for example, because they cannot fold correctly) could be made, but rapidly removed. To arrive at a true set of protein-encoding genes, we cannot rely on computational techniques alone, but must continue to characterize proteins and their functions.

These problems provide scope for estimates of human gene number to vary widely. Although recent estimates are converging in the 30,000–40,000 range (as opposed to earlier estimates of 100,000 or so), it could be many years before we have the final answer. **P.B. & R.C.**

single nucleotide polymorphisms — the base differences from human to human — the draft also provides a framework for understanding the genetic basis and evolution of many human characteristics.

With the draft in hand, researchers have a new tool for studying the regulatory regions and networks of genes. Comparisons with other genomes should reveal common regulatory elements, and the environments of

genes shared with other species may offer insight into function and regulation beyond the level of individual genes. The draft is also a starting point for studies of the three-dimensional packing of the genome into a cell's nucleus. Such packing is likely to influence gene regulation.

On a more applied note, the information can be used to exploit technologies such as chips made using DNA or proteins, complementing more traditional approaches. Such chips could now, for instance, contain all the members of a protein family, making it possible to find out which are active in particular diseased tissues. A new world of biotechnology will provide tools and information by exploiting genome data.

Sequencing the tough leftovers of the human genome will be essential. Without a finished sequence, we will not know what we are missing. Each missed gene is potentially a missed drug target, and even gene-poor areas might be critical for gene regulation. Nevertheless, we must now confront the fact that the era of rapid growth in human genomic information is over. The challenge we face is nothing less than understanding

how this comparatively small set of genes creates the diversity of phenomena and characteristics that we see in human life. The human genome lies before us, ready for interpretation.

Peer Bork and Richard Copley are at EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany.

Peer Bork is at the Max-Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse 10, 13125 Berlin-Buch, Germany.

e-mails: Peer.Bork@EMBL-Heidelberg.de

Richard.Copley@EMBL-Heidelberg.de

1. International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
2. Venter, J. C. et al. *Science* **291**, 1304–1351 (2001).
3. Dunham, I. et al. *Nature* **402**, 489–495 (1999).
4. The Chromosome 21 Mapping and Sequencing Consortium *Nature* **405**, 311–319 (2000).
5. The C. elegans Sequencing Consortium *Science* **282**, 2012–2018 (1998).
6. Collins, F. S. et al. *Science* **282**, 682–689 (1998).
7. Liang, F. et al. *Nature Genet.* **25**, 239–240 (2000).
8. Liang, F. et al. *Nature Genet.* **26**, 501 (2000).
9. Shoemaker, D. D. et al. *Nature* **409**, 922–927 (2001).
10. The Arabidopsis Sequencing Consortium *Cell* **100**, 377–386 (2000).
11. Adams, M. D. et al. *Science* **287**, 2185–2195 (2000).
12. Normile, D. & Pennisi, E. *Science* **285**, 2038–2039 (1999).
13. Aparicio, S. *Nature Genet.* **25**, 129–130 (2000).
14. Goffeau, A. et al. *Nature* **387** (suppl.), 1–105 (1997).
15. The Arabidopsis Genome Initiative *Nature* **408**, 796–815 (2000).

are predicted to exist, 60% have some sequence similarity to proteins from other species whose genomes have been sequenced. Just over 40% of the predicted human proteins share similarity with fruitfly or worm proteins. And 61% of fruitfly proteins, 43% of worm proteins and 46% of yeast proteins have sequence similarities to predicted human proteins.

But what about the proteins whose sequences show no strong similarity to known proteins from other species? Over a third of the yeast, fruitfly, worm and human proteins fall into this class. These proteins might retain similar functions, even though their sequences have diverged. Or they might have acquired species-specific functions.

Alternatively, we may need to entertain the possibility that the open reading frames that encode these proteins are maintained in a new way, one that is independent of the precise amino-acid sequence and thus is free to evolve rapidly. (An open reading frame is the part of a gene encoding the amino-acid sequence of its protein product.) After all, we know that cells have at least one mechanism, called nonsense-mediated decay of mRNA, for detecting imperfect open reading frames irrespective of the amino-acid sequence that they encode⁸.

It will be interesting to see the extent to which the number of human proteins in this rapidly evolving class decreases as the genomes of other vertebrates, such as mice, are sequenced. This will give us an indication of just how fast these proteins are changing. Indeed, there is already evidence from studies of flies⁹ and worms¹⁰ that these rapidly evolving proteins are less likely to have essential functions, consistent with their being less likely to be conserved during evolution.

Such comparisons of distantly related genomes are fascinating from an evolutionary point of view. But comparison of closely related genomes will be much more important in addressing the key problem now facing genomics — determining the function of individual DNA segments. The concept is simple: segments that have a function are more likely to retain their sequence during evolution than non-functional segments. So DNA segments that are conserved between species are likely to have important functions. The ideal species for comparison are those whose form, physiology and behaviour are as similar as possible, but whose genomes have evolved sufficiently that non-functional sequences have had time to diverge. In practice, there may be no one ideal species, because different genes and regulatory sites evolve at different rates. Nevertheless, this approach has a long history of success, and becomes progressively more efficient as the cost of DNA sequencing declines.

One use of such sequence comparisons is

The draft sequences

Comparing species

Gerald M. Rubin

Comparing the human genome sequences with those of other species will not only reveal what makes us genetically different. It may also help us understand what our genes do.

How are the differences between humans and other organisms reflected in our genomes? How similar are the numbers and types of proteins in humans, fruitflies, worms, plants and yeast? And what does all of this tell us about what makes a species unique? With the publication of the draft human genome sequences, on page 860 of this issue¹ and in this week's *Science*², we can start to compare the sequences of vertebrate, invertebrate and plant genomes in an attempt to answer these questions.

An obvious place to start our comparison is the total number of genes in each species. Here is a real surprise: the human genome probably contains between 25,000 and 40,000 genes, only about twice the number needed to make a fruitfly³, worm⁴ or plant⁵. We know that there is a higher degree of 'alternative splicing' in humans than in other species. In other words, there are often many more ways in which a gene's protein-coding sections (exons) can be joined together to create a functional messenger RNA molecule, ready to be translated into protein. So more proteins are encoded per gene in humans than in other species.

Even so, we cannot escape the conclusion

— drawn previously from comparisons of simpler genomes⁶ — that physical and behavioural differences between species are not related in any simple way to gene number. Many researchers, struck by the fact that there are four times as many genes in some gene families in the human genome compared with fruitflies⁷, extrapolated from these cases and suggested that the human genome might be the product of two doublings of the whole of a simpler genome found in the common ancestor of fruitflies and humans. But, as the analyses of the human genome show^{1,2}, if such doublings did occur, the evidence for them has since been obscured by massive gene loss and amplification of particular gene families in the human genome.

Individual proteins often feature discrete structural units, called domains, that are conserved in evolution. More than 90% of the domains that can be identified in human proteins are also present in fruitfly and worm proteins, although they have been shuffled to create nearly twice as many different arrangements in humans^{1,2}. Thus, vertebrate evolution has required the invention of few new domains. Of the human proteins that

to determine the structure of genes — which parts (the exons) make their way into a functional mRNA molecule and which do not (the introns). The high degree of alternative splicing in vertebrates makes this comparative approach particularly important. Gene-finding computational algorithms cannot easily predict the existence of alternative forms of an mRNA without experimental information, but this information is difficult to come by in the case of rare mRNAs. For example, an exon that is used in only a few cells of the human brain might never be experimentally detected in an mRNA. But that exon's sequence would probably be conserved in the mouse genome.

Comparing the genomes of closely related species can also help in identifying gene-control regions. This approach has been used for over two decades¹¹, and has been validated by showing that the conserved sequences indeed correspond to functional control elements in individual genes¹². But this computational problem is more difficult than identifying exons, and it will be challenging to scale up to a genome-wide level. The proteins that control gene expression by recognizing regulatory regions often detect sequence features that elude the best computer algorithms, and may use information from contacts with other proteins that is difficult to model. Proteins are simply cleverer than computers.

That said, our knowledge of the DNA-

binding properties of individual proteins, as well as the structural features of the DNA sites to which they bind, continues to increase. Moreover, we can use experimental evidence; for example, genes that are expressed together might be expected to share control elements. And, as methods for comparing sequences continue to improve, we can expect to learn more about elusive features of the genome, such as genes encoding RNAs that do not encode proteins¹³, start points of DNA replication, and genetic elements that control chromosome structure. ■

Gerald M. Rubin is in the Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94708-3200, and the Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, Maryland 20815-6789, USA.

e-mail: gerry@fruitfly.berkeley.edu

1. International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* *Science* **291**, 1304–1351 (2001).
3. Adams, M. D. *et al.* *Science* **287**, 2185–2195 (2000).
4. The *C. elegans* Sequencing Consortium *Science* **282**, 2012–2018 (1998).
5. The Arabidopsis Genome Initiative *Nature* **408**, 796–815 (2000).
6. Rubin, G. M. *et al.* *Science* **287**, 2204–2215 (2000).
7. Spring, J. *FEBS Lett.* **400**, 2–8 (1997).
8. Hentze, M. W. & Kulozik, A. E. *Cell* **96**, 307–310 (1999).
9. Ashburner, M. *et al.* *Genetics* **153**, 179–219 (1999).
10. Fraser, A. G. *et al.* *Nature* **408**, 325–330 (2000).
11. Ravetch, J. V., Kirsch, I. R. & Leder, P. *Proc. Natl Acad. Sci. USA* **77**, 6734–6738 (1980).
12. Fortini, M. E. & Rubin, G. M. *Genes Dev.* **4**, 444–463 (1990).
13. Lee, R. C., Feinbaum, R. L. & Ambros, V. *Cell* **75**, 843–854 (1993).

Single nucleotide polymorphisms

From the evolutionary past...

Mark Stoneking

Single nucleotide polymorphisms are the bread-and-butter of DNA sequence variation. They provide a rich source of information about the evolutionary history of human populations.

Studies of genetic variation in human populations began inauspiciously¹. The first such study — of ABO blood-group frequencies — was carried out by two Polish immunologists, Ludwik and Hanka Hirsfeld, at the end of the First World War. This work was notable for its broad coverage of the world's populations, large sample sizes and scrupulous attention to anthropological details. Yet the Hirsfelds still ran into difficulties in publishing in *The Lancet*, the premier medical journal of the time. The editor could not see the relevance of their work, and so this seminal study of human genetic variation first appeared in an obscure anthropological journal². The relevance became abundantly clear when Felix Bernstein subsequently used the Hirsfelds' data to demonstrate that the ABO blood-group frequencies were better explained by a single gene with three variants (alleles), and not —

as prevailing wisdom then held — two genes each with two alleles³.

Happily, times have changed, diversity is now all the rage^{4,5}, and editors have become more appreciative of the importance of human genetic variation. The latest evidence of that is the paper on page 928 of this issue⁶, which reports the identification and mapping of 1.4 million single nucleotide polymorphisms (SNPs, pronounced 'snips') in the human genome. The paper is the result of the labours of a large collaboration, The International SNP Map Working Group.

So, what are SNPs? Quite simply, they are the bread-and-butter of DNA sequence variation — polymorphism, to those in the business. A DNA sequence is a linear combination of four nucleotides; compare two sequences, position by position, and wherever you come across different nucleotides at the same position, that's a SNP (see Fig. 1 on

page 823). So SNPs reflect past mutations that were mostly (but not exclusively) unique events, and two individuals sharing a variant allele are thereby marked with a common evolutionary heritage. In other words, our genes have ancestors, and analysing shared patterns of SNP variation can identify them.

However, the real importance of SNPs is that there are so many of them. One estimate⁷ is that comparing two human DNA sequences results in a SNP every 1,000–2,000 nucleotides. That may not sound like much until you realize that there are 3.2 billion nucleotides in the human genome, which translates into 1.6 million–3.2 million SNPs. And that's just from comparing two sequences — the total number of SNPs in humans is obviously much more. Most human variation that is influenced by genes can be traced to SNPs, especially in such medically (and commercially) important traits as how likely you are to become afflicted with a particular disease, or how you might respond to a particular pharmaceutical treatment, as discussed by Chakravarti⁸ on the following page. And even when a SNP is not directly responsible, the sheer number of SNPs means they can also be used to locate genes that influence such traits.

The deluge of SNPs reported by the SNP working group⁶ also promises great things for those of us who analyse patterns of molecular genetic variation to reconstruct the evolutionary history of human populations. Our genes contain the signature of an expansion from Africa within the past 150,000 years or so⁹. But there is still debate as to whether the modern humans from Africa completely replaced archaic non-African populations with no interbreeding, or whether we perhaps carry the vestiges of Neanderthal or other archaic non-African genes.

Demonstrating a recent African origin for every single one of our 3.2 billion nucleotides goes beyond the bounds of reason or necessity, but there is still much to be learned. For a start, most of our insights into molecular anthropology arise from DNA in mitochondria and (more recently) polymorphisms of the Y chromosome. This is because these DNA sequences are haploid — that is, represented just once in each cell, in contrast to the other chromosomes, which are represented twice — and they are inherited from just one parent, so they do not undergo the usual sequence shuffling (recombination) during egg and sperm production. This makes them easier to analyse and extremely informative. But both suffer from the drawback that, in the absence of recombination, they behave as single genes, and the history of any single gene can differ from that of a population or species because of natural selection or chance events involving that gene.

Accurate inferences concerning popula-

tion history demand the analysis of several genes, with the most promising approach involving haplotypes¹⁰, which consist of several closely spaced (linked) polymorphisms. The advantage of haplotypes over simply analysing polymorphisms at random is that there is valuable information in the associations between linked polymorphisms — the whole is greater than the sum of the parts. So the 1.4 million SNPs are a welcome resource that will greatly help in identifying haplotypes for tracing human evolutionary history, especially those that might reveal archaic non-African ancestry.

However, answering all of our questions about human evolutionary history will not be as simple as mining the SNP database and determining haplotypes in a representative sample of worldwide populations. There are four main reasons for that.

First, to be really useful, the SNPs in the database should really be SNPs, and not errors or artefacts, and they should be polymorphic in other samples, not just the sample of individuals used to find the SNPs. An important aspect of the SNP working group's data is that 1,585 SNPs were chosen for further verification, of which about 95% turned out to be true SNPs, which is good news indeed. Moreover, 1,276 SNPs were tested on additional population samples and at least 82% were polymorphic, which is reassuring.

Second, one might ask why only 0.1% of the 1.4 million SNPs were verified and tested. The answer is that our ability to determine allele frequencies efficiently and inexpensively for large numbers of SNPs lags behind our ability to simply identify them. This situation is reminiscent of the beginnings of the Human Genome Project, when developing technology was a primary concern and it was not at all clear how the 3.2 billion nucleotides were going to be determined. But human ingenuity won out then, and given the number of bright and capable minds now wrestling with the SNP-typing problem, one or more solutions should soon be at hand (especially with the motivation of lucrative commercial applications).

Third, a problem known as ascertainment bias can complicate the interpretation of results based on SNPs. For example, SNPs that were found to be polymorphic in European populations will overestimate genetic diversity in European as opposed to non-European populations. Moreover, the probability of finding a SNP, and the frequency of polymorphism at a SNP, depends on how many times a particular DNA segment was sequenced, and from how many individuals. The SNP working group report some intriguing preliminary findings regarding how SNP diversity is apportioned among chromosomes. But further work is required to see if these are truly biological differences, or if they instead reflect

ascertainment biases. Ascertainment bias is not an insurmountable problem — statistical geneticists love this sort of challenge and are already coming up with creative solutions¹¹. Even so, SNP-finders must keep careful track of how their SNPs were ascertained.

Fourth, the emphasis in the SNP database is on SNPs where both of the alleles occur at high frequency, because these will be most useful for disease-association studies. In general, the higher the frequency of a SNP allele, the older the mutation that produced it, so high-frequency SNPs largely predate human population diversification. But many questions in human evolution involve specific migrations (such as the colonization of Polynesia or the Americas) for which population-specific alleles are most informative — indeed, this is one of the attractions of mitochondrial-DNA and Y-chromosome analyses for such questions, because population-specific alleles can be readily found. It is unlikely that Polynesian-specific SNPs are present in the database, so more work will be required to find such informative, population-specific SNPs.

Still, one can imagine that in the not-too-distant future the details of human population history will have been fleshed out, at least to the extent possible by analysing genetic variation in extant populations. What then? One area that is receiving increasing attention is the detection of the effects of natural selection in human populations¹². Using SNPs to find chromosomal regions with abnormally low levels of varia-

tion is a particularly promising way of detecting the genomic signature of selection for favourable mutations¹³.

Another area of increasing interest is identifying the molecular genetic basis of 'normal' phenotypic variation⁴ — that is, variation of the old-fashioned, morphological kind, which is a traditional concern of anthropology. Molecular anthropology has for the most part concentrated on the molecules and what their diversity tells us about human evolution. With the advent of the human genome sequence and the SNP database, the ultimate in molecular tools, we are ironically now poised to focus on phenotypes and what their diversity tells us about human evolution — thereby bringing the anthropology back into molecular anthropology.

Mark Stoneking is at the Max Planck Institute for Evolutionary Anthropology, Inselstrasse 22, D-04103 Leipzig, Germany.

e-mail: stoneking@eva.mpg.de

1. Mourant, A. E. *Blood Relations* p.13 (Oxford Univ. Press, 1983).
2. Hirschfeld, L. & Hirschfeld, H. *Anthropologie* **29**, 505–537 (1919).
3. Crow, J. F. *Genetics* **133**, 4–7 (1993).
4. Weiss, K. M. *Genome Res.* **8**, 691–697 (1998).
5. Collins, F. S., Brooks, L. D. & Chakravarti, A. *Genome Res.* **8**, 1229–1231 (1998).
6. The International SNP Map Working Group *Nature* **409**, 928–933 (2001).
7. Li, W. H. & Sadler, L. A. *Genetics* **129**, 513–523 (1991).
8. Chakravarti, A. *Nature* **409**, 822–823 (2001).
9. Stoneking, M. *Evol. Anthropol.* **2**, 60–73 (1993).
10. Tishkoff, S. A. *et al. Science* **271**, 1380–1387 (1996).
11. Kuhner, M. K., Beerli, P., Yamato, J. & Felsenstein, J. *Genetics* **156**, 439–447 (2000).
12. Przeworski, M., Hudson, R. R. & Di Rienzo, A. *Trends Genet.* **16**, 296–302 (2000).
13. Nurminsky, D., De Aguiar, D., Bustamante, C. D. & Hartl, D. L. *Science* **291**, 128–130 (2001).

Single nucleotide polymorphisms

...to a future of genetic medicine

Aravinda Chakravarti

Single base differences between human genomes underlie differences in susceptibility to, or protection from, a host of diseases. Hence the great potential of such information in medicine.

The beginning of the Human Genome Project, over a decade ago, was accompanied by a cantankerous debate over whose genome was to be sequenced. Would it be a single individual? A celebrity, perhaps (widely rumoured to be Jim Watson, co-discoverer of the structure of DNA)? Or would several genomes, from many individuals, be studied? The discussion struck at the very heart of genetics. As the study of inherited variation between individuals, genetics might not immediately benefit from the sequence of a single genome. But even one genome would be immensely revealing to the science of deciphering the molecular blueprint of a species. Fortunately, geneticists were not forced to make this choice. Papers in this issue describe not only a single,

history-making human genome sequence, composed of little bits from many humans¹ (page 860), but also some 1.4 million sites of variation mapped along that reference sequence² (page 928).

But why this preoccupation with sequence variation, with the fact that no two humans (except identical twins) are genetically the same? The answer is that such variations, or 'polymorphisms', are markers of genes and genomes with which researchers perform genetic analysis in an outbred species where matings cannot be controlled. The fields of human and medical genetics simply cannot exist without understanding this variation.

It has become clear that the two 'genomes' that each of us carry, inherited

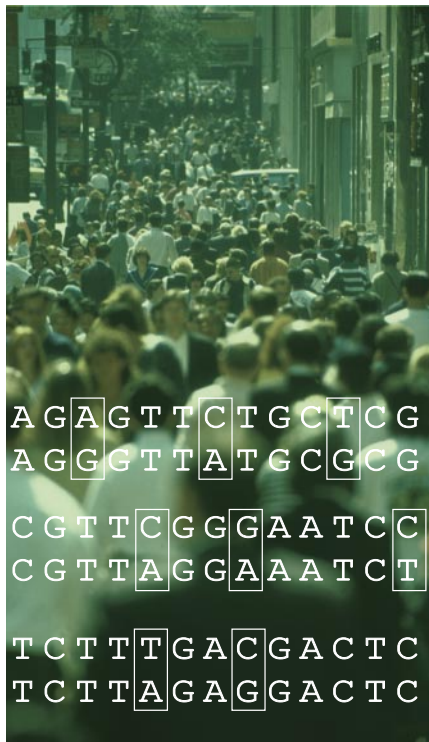


Figure 1 The most common sources of variation between humans are single nucleotide polymorphisms (SNPs) — single base differences between genome sequences. Fragments of two sequences, with eight SNPs, are shown.

from our parents, most often differ — from each other, and from the genomes of other humans — in terms of single base changes¹ (Fig. 1). The twentieth century saw the identification of only a few thousand of these so-called single nucleotide polymorphisms (SNPs, or ‘snips’ to the streetwise). In just the first year of the new century, this number has been increased one-thousand-fold². Beyond the numbers, the excitement today comes from precise knowledge of where these sites of variation are in the genome². The 1.42 million known SNPs are found at a density of one SNP per 1.91 kilobases. This means that more than 90% of any stretches of sequence 20 kilobases long will contain one or more SNPs. The density is even higher in regions containing genes. The International SNP Map Working Group² estimates that they have identified 60,000 SNPs within genes (‘coding’ SNPs), or one coding SNP per 1.08 kilobases of gene sequence. Moreover, 93% of genes contain a SNP, and 98% are within 5 kilobases of a SNP. For the first time, nearly every human gene and genomic region is marked by a sequence variation.

These data provide interesting first glimpses into the pattern of variation across the genome. Variation is commonly assessed by nucleotide diversity — the number of base differences between two genomes, divided by the number of base pairs

compared. Nucleotide diversity is a sensitive indicator of biological and historical factors that have affected the human genome³. The nucleotide diversity in gene-containing regions has been estimated to be 8 differences per 10 kilobases^{4,5}; we now know that the genome-wide average is similar, 7.51 differences per 10 kilobases (ref. 2). The variation between individual non-sex chromosomes is small, and lies in the range 5.19 (for chromosome 21) to 8.79 (for chromosome 15) differences per 10 kilobases (ref. 2).

Strikingly, humans vary least in their sex chromosomes. The variation between different X chromosomes is about 4.69 differences per 10 kilobases, and it is very much lower for the Y chromosome (1.51 differences per 10 kilobases). This is because the sex chromosomes have patterns of mutation and recombination (the swapping of similar DNA segments during the generation of eggs and sperm) that differ both from each other and from the non-sex chromosomes. Moreover, fewer ancestors have contributed to the sex chromosomes, which are therefore less variable than the non-sex chromosomes.

Perhaps not surprisingly, some genomic regions have significantly lower or higher diversity than the average. For example, the HLA locus, which encodes proteins that present antigens to the immune system, shows the greatest diversity. Such comparisons within genomes will be essential to our understanding of how variation shapes biochemical and cellular functions, and in illuminating past human evolution, as discussed in ref. 3, and by Stoneking in the preceding article (page 821; ref. 6).

But the main use of the human SNP map will be in dissecting the contributions of individual genes to diseases that have a complex, multigene basis. Knowledge of genetic variation already affects patient care to some degree. For example, gene variants lead to tissue and organ incompatibility, affecting the success of transplants. And the mainstay of medical genetics has been the study of the rare gene variants that lie behind inherited diseases such as cystic fibrosis.

But variations in genome sequences underlie differences in our susceptibility to, or protection from, all kinds of diseases; in the age of onset and severity of illness; and in the way our bodies respond to treatment. For example, we already know that single base differences in the *APOE* gene are associated with Alzheimer’s disease, and that a simple deletion within the chemokine-receptor gene *CCR5* leads to resistance to HIV and AIDS. The benefit of the SNP map is that it covers the entire genome. So, by comparing patterns and frequencies of SNPs in patients and controls, researchers can identify which SNPs are associated with which diseases^{7–9}. Such research will bring about ‘genetic medicine’, in which knowledge of our uniqueness

will alter all aspects of medicine, perceptibly and forever.

Studies of SNPs and diseases will become more efficient when a few more problems are solved³. First, although 82% of SNP variants are found at a frequency of more than 10% in the global human population, the ‘micro-distribution’ of SNPs in individual populations is not known. Second, not all SNPs are created equal, and it will be essential to know as much as possible about their effects from computational analyses before studying their involvement in disease. For example, each SNP can be classified by whether it is coding or not. Coding SNPs can be classified by whether they alter the sequence of the protein encoded by the altered gene. Changes that alter protein sequences can be classified by their effects on protein structure. And non-coding SNPs can be classified according to whether they are found in gene-regulating segments of the genome¹⁰ — many complex diseases may arise from quantitative, rather than qualitative, differences in gene products. Third, the technology for assaying thousands of SNPs, in thousands of patients and controls⁷, is not yet fully developed, although there are some creative ideas around.

In the twentieth century, humans were not the geneticists’ species of choice. The emphasis then was on understanding gene structure and function. Now, geneticists will concentrate increasingly on understanding physical and behavioural characteristics. Here, our species, with its obsession with self-examination, will make a superior subject. We will also see more studies of how natural variation leads to each one of our qualities. To some, there is a danger of genomania, with all differences (or similarities, for that matter) being laid at the altar of genetics¹¹. But I hope this does not happen. Genes and genomes do not act in a vacuum, and the environment is equally important in human biology. By identifying variation across the whole genome, the SNP map² may be our best route yet to a better understanding of the roles of nature and (not versus) nurture. ■

Aravinda Chakravarti is at the McKusick–Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, 600 North Wolfe Street, Jefferson Street Building 2-109, Baltimore, Maryland 21287, USA.

e-mail: aravinda@jhmi.edu

1. International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
2. The International SNP Map Working Group *Nature* **409**, 928–933 (2001).
3. Chakravarti, A. *Nature Genet.* **21** (suppl.), 56–60 (1999).
4. Halushka, M. K. et al. *Nature Genet.* **22**, 239–247 (1999).
5. Cargill, M. et al. *Nature Genet.* **22**, 231–238 (1999).
6. Stoneking, M. *Nature* **409**, 821–822 (2001).
7. Risch, N. & Merikangas, K. *Science* **273**, 1516–1517 (1996).
8. Lander, E. S. *Science* **274**, 536–539 (1996).
9. Collins, F. S., Guyer, M. S. & Chakravarti, A. *Science* **278**, 1580–1581 (1997).
10. Loo, G. G. et al. *Science* **288**, 136–140 (1999).
11. Lewontin, R. *It Ain’t Necessarily So: The Dream of the Human Genome and Other Illusions* (New York Review of Books, 2000).

Guide to the draft human genome

Tyra G. Wolfsberg*, Johanna McEntyre† & Gregory D. Schuler†

* Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

† National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

There are a number of ways to investigate the structure, function and evolution of the human genome. These include examining the morphology of normal and abnormal chromosomes, constructing maps of genomic landmarks, following the genetic transmission of phenotypes and DNA sequence variations, and characterizing thousands of individual genes. To this list we can now add the elucidation of the genomic DNA sequence, albeit at 'working draft' accuracy. The current challenge is to weave together these disparate types of data to produce the information infrastructure needed to support the next generation of biomedical research. Here we provide an overview of the different sources of information about the human genome and how modern information technology, in particular the internet, allows us to link them together.

The ultimate goal of the Human Genome Project is to produce a single continuous sequence for each of the 24 human chromosomes and to delineate the positions of all genes. The working draft sequence described by the International Human Genome Sequencing Consortium was constructed by melding together sequence segments derived from over 20,000 large-insert clones¹. All of the results of this analysis are available on a web site maintained by the University of California at Santa Cruz (<http://genome.ucsc.edu>). Over the next few years, draft quality sequence will be steadily replaced by more accurate data. The National Center for Biotechnology Information (NCBI) has developed a system for rapidly regenerating the genomic sequence and gene annotation as sequences of the underlying clones are revised (<http://www.ncbi.nlm.nih.gov/genome/guide>). Undoubtedly, others will apply a variety of approaches to large-scale annotation of genes and other features. One such project is Ensembl, a joint project of the European Bioinformatics Institute (EBI) and the Sanger Centre (<http://www.ensembl.org>).

Pinpointing new genes

Recent estimates have placed the number of human genes at 25,000–35,000 (refs 2, 3). More than 10,000 human genes have been catalogued in the Online Mendelian Inheritance in Man⁴ (OMIM), which documents all inherited human diseases and their causal gene mutations. Integration of the information contained in OMIM with the working draft is facilitated by the fact that it has already been tied to reference messenger RNA sequences through a collaborative effort between OMIM, the Human Gene Nomenclature Committee and the NCBI^{5,6}. As a result, the positions of many known genes have been determined by alignment of mRNAs with genomic sequences. For the remaining genes, we must currently resort to computational gene-finding methods (reviewed in refs 7, 8).

When mRNA species align differently to a genomic sequence, this indicates that alternative splicing has taken place. In the current set of full-length reference mRNAs, 11,174 transcripts have been sequenced from 10,742 distinct genes (2.4% of the genes have multiple splicing variants). Alignments of expressed sequence tag (EST) sequences to the working draft sequence, however, suggest that about 60% of human genes have multiple splicing variants, which has important implications for the complexity of human gene expression¹. By their sheer numbers (currently over 2.5 million) we might expect ESTs to sample a larger fraction of splicing variants than would be the case for more traditional targeted approaches. For example, alignment of the mRNA of the membrane-bound metalloprotease-disintegrin ADAM23 (ref. 9) to

the draft genome reveals that the gene consists of at least 23 exons. Of the many ESTs that also align to the ADAM23 locus, one lacks the exon that encodes the transmembrane domain, which suggests an alternatively spliced, soluble protein. Although this is a biologically plausible conclusion, one should exercise caution when interpreting such results: ESTs are partial single-pass sequences that have been associated with a variety of artefacts, including sequencing errors and improper splicing^{10,11}.

Finding relatives

Genes can be found through an implied relationship to something else—for example, being a putative orthologue (related to a gene in another species). To do this, it is useful to search the genomic sequence or, preferably, its mRNA sequences and protein products, using BLAST¹². As an example, we use the mouse *Lmx1b* gene, which encodes a LIM homeobox protein that is important in pattern development¹³. When we used the protein sequence encoded by *Lmx1b* as a query in a BLAST search against the working draft human genome sequence, the best match was to a region of 9q34, and the positions of the alignments line up with the exons of the human *LMX1B* gene (Fig. 1a).

Additional support for two genes being orthologous comes from the mouse–human homology map. Despite being separated by 200 million years of evolution, mouse and human genes often fall into homologous chromosomal regions that share a conserved gene order (synteny). In fact, the working draft sequence has helped to refine the homology map and provides inferred map positions for many mouse genes¹. The two homeobox genes fall within a conserved syntenic block between mouse chromosome 2 and human chromosome 9 (Fig. 1b). Furthermore, the human *LMX1B* gene has been implicated in nail patella syndrome (NPS), an autosomal recessive disorder characterized by limb and kidney defects. A mouse in which *Lmx1b* has been inactivated shows a phenotype that is strikingly similar to NPS¹³. Besides providing additional support for the conclusion of orthology, this connection may provide a useful mouse model for the human disorder. In this way, information from OMIM and mouse mutants can further define human genes.

Another way to find a gene is by looking for paralogues—family members derived by gene duplication. As an example, we used human ADAM23, which maps to 2q33 (ref. 14). In a BLAST search against a set of proteins predicted from the draft sequence, aside from matching itself, the best match was to a peptide from chromosome 20. No ADAM family member has previously been mapped to this chromosome. The predicted protein encoded by this gene does not begin with a methionine and appears to be incomplete at its amino terminus when aligned with other family

members: this could be due to an erroneous protein prediction or a gap in the draft sequence. Computational analysis can reveal protein family domains and their relationships to three-dimensional protein structures. In this case, the putative ADAM paralogue contains both the zinc metalloprotease and disintegrin motifs characteristic of the ADAM family. Critical amino acids of the metalloprotease domain are conserved in the putative paralogue (Fig. 2b), including a trio of histidine residues in the active site (shaded yellow), which are important in complexing the zinc ion (Fig. 2a). The finding that the sequence of the new predicted ADAM member has an intact active site suggests that the predicted gene is functional, rather than being a pseudogene.

Searching by position

It is sometimes desirable to find genes by their position in the

genome, rather than by sequence similarity. For example, when genetic or cytogenetic analysis has implicated a particular region in the aetiology of a disease, it is of interest to see what genes lie in the region. A natural way to describe positions in a sequence would be by base coordinates, but this is impractical for the working draft sequence, as the sequence is still being revised. Cytogenetic band nomenclature is more commonly used to describe positions in the genome, and many human diseases are linked to chromosomal deletions, amplifications and translocations¹⁵. However, to be useful in conjunction with the working draft, these designations must be related to the sequence. Towards this end, a consortium has integrated this information using fluorescence *in situ* hybridization (FISH) to localize BAC clones that also bear sequence tags that can be found in the draft sequence¹⁶. In addition to providing cytogenetic coordinates as entry points into the genome, they also provide mapped clone reagents that may be useful in further experimental work.

Another way to describe positions in the genome is relative to mapped sequence tagged site (STS) markers. This is particularly useful in positional cloning projects, where candidate regions are usually defined by polymorphic STSs used in genetic linkage analysis. For example, the breast cancer susceptibility locus *BRCA2* was originally localized by fine genetic mapping to a 600-kilobase (kb) interval on chromosome 13 centred around the STS marker D13S171 (ref. 17). STS markers from several genetic and physical maps have been localized in the working draft sequence using a procedure known as electronic PCR¹⁸. Thus, by simply looking up the position of D13S171, we can see the region around what is now known to be *BRCA2*, together with other features such as adjacent genes and markers, translocation breakpoints, and genetic variations.

Variations on a theme

A map of DNA sequence variations will aid our understanding of complex diseases and human population dynamics. The most common class of variation is the single nucleotide polymorphism (SNP) and the total number of SNPs in the public database (dbSNP)¹⁹ now exceeds 2.5 million, representing 1.5 million unique SNP loci. Because database entries include flanking sequence surrounding the polymorphic base(s), it is possible to localize variations within the working draft by simple sequence alignment²⁰.

Histone deacetylase 3 (HDAC3) is a nucleosome-remodelling enzyme that deacetylates the lysine residues of histones, affecting transcriptional repression²¹. Its genomic region contains seven mapped SNPs near the locus, one of which falls within the coding region—a G-to-C substitution that results in the nonsynonymous substitution Arg265Pro in the protein product. The three-dimensional structure of an *Aquifex aeolicus* homologue shows that this residue is at the lip of the active-site pocket²². Of the two classes of eukaryotic histone deacetylase, one most often has Arg at this position and the other most often has Pro²³, a trait shared with the bacterial members of the histone deacetylase superfamily. This SNP might therefore occur at a functionally interesting site, and also gives pause for speculation: as bacterial members of the superfamily predominantly have a Pro in this position, perhaps Pro is the ancient residue at this site, and not Arg. Note that several high-throughput SNP discovery methods have been used to generate these data and not all SNPs have been rigorously validated.

Conclusions

The draft sequence provides us with the first comprehensive integration of diverse genomic resources. The mapping of ESTs, gene predictions, STSs and SNPs onto the draft sequence can enable identification of alternative splicing, orthologues, paralogues, map positions and coding sequence variations. Users should remember,

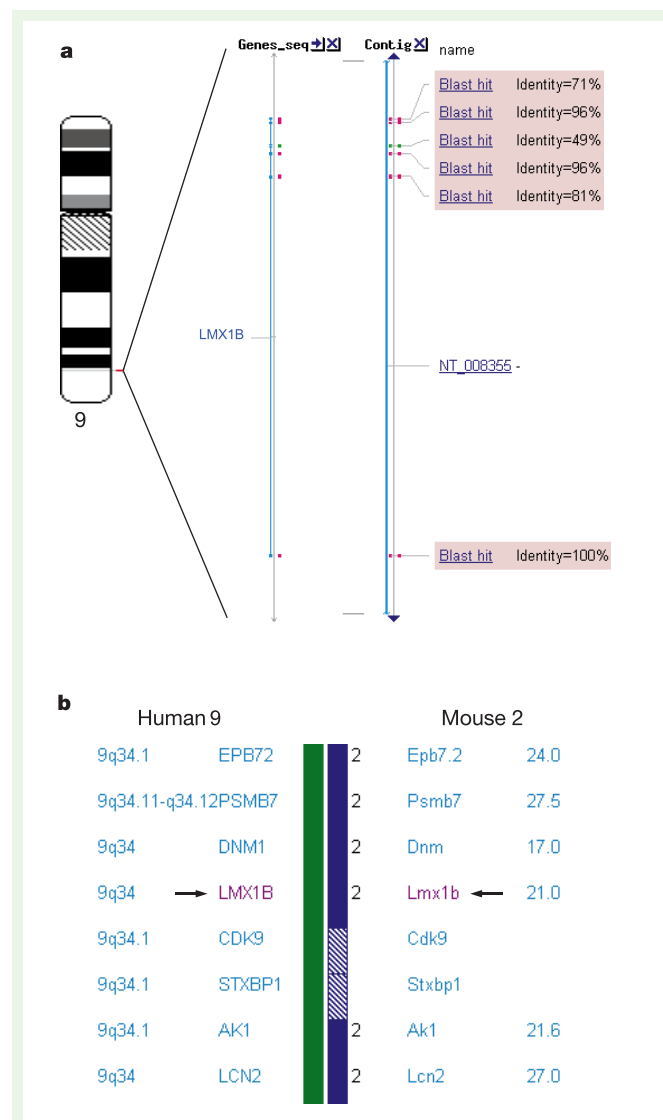


Figure 1 *Lmx1b* encodes a transcription factor that helps to control the trajectory of motor axons during mammalian limb development. **a**, The results of a search of the six-frame translation of the draft genome sequence on the NCBI site using mouse *Lmx1b* protein NP_034855.1 as a query and using TBLASTN with standard search parameters. The best match was to a region of chromosome 9 that contains the human *LMX1B* gene. **b**, The mouse *Lmx1b* and the human *LMX1B* genes lie within a conserved syntenic block of genes, in mouse on chromosome 2 and in human on chromosome 9. This conservation of gene order supports the theory that *Lmx1b* and *LMX1B* are orthologous.

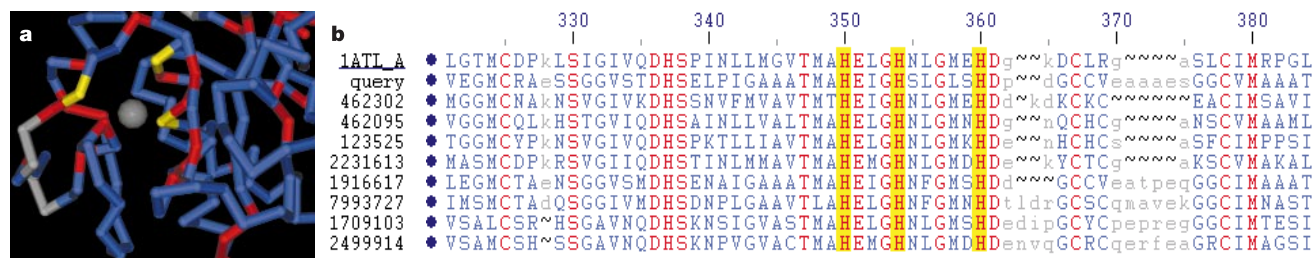


Figure 2 The ADAM23 protein sequence NP_003803.1 was used in a BLASTP search of the Ensembl confirmed peptides produced from the 5 Sep 2000 version of the working draft sequence. As of October 2000, the best match was to the peptide ENSP00000025626 derived from chromosome 2, the predicted peptide for ADAM23. The second match was to peptide ENSP00000072108, derived from chromosome 20 and falling within GenBank Acc. No. AC055771.2. We used this predicted peptide to search a database of Pfam²⁴ and SMART²⁵ protein domains that are aligned with protein structures. This Conserved Domain Database (CDD) search at NCBI resulted in hits to reprolysin and disintegrin domains. **a**, Pfam family 01421, Reprolysin, has a structure associated with it: the zinc-dependent metalloprotease Atrolysin C (PDB: 1ATL). **b**, The query ENSP00000072108 aligns with the 1ATL protein sequence and eight other ADAMs from the Pfam Reprolysin entry. In the structure and alignment, red indicates conserved residues; grey indicates non-aligned sequences. The three histidines of the metalloprotease active site, which complex with the zinc ion, are highlighted in yellow. The structure and alignment were created using the structure viewing program Cn3D.

though, that these genomic resources represent a work-in-progress, and will evolve as the genome is finished and computation methods further refined.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25**, 232–234 (2000).
3. Roest Crolius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25**, 235–238 (2000).
4. McKusick, V. A. *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders* (Johns Hopkins Univ. Press, Baltimore, 1998).
5. Maglott, D. R., Katz, K. S., Sicotte, H. & Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* **28**, 126–128 (2000).
6. Pruitt, K. D., Katz, K. S., Sicotte, H. & Maglott, D. R. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* **16**, 44–47 (2000).
7. Guigo, R., Agarwal, P., Abril, J. F., Burset, M. & Fickett, J. W. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* **10**, 1631–1642 (2000).
8. Stormo, G. D. Gene-finding approaches for eukaryotes. *Genome Res.* **10**, 394–397 (2000).
9. Sagane, K., Ohya, Y., Hasegawa, Y. & Tanaka, I. Metalloproteinase-like, disintegrin-like, cysteine-rich proteins MDC2 and MDC3: novel human cellular disintegrins highly expressed in the brain. *Biochem. J.* **334**, 93–98 (1998).
10. Wolfsberg, T. G. & Landsman, D. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* **25**, 1626–1632 (1997).
11. Wolfsberg, T. G. & Landsman, D. in *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (eds Baxevanis, A. D. & Ouellette, B. F. F.) (Wiley-Liss, Inc., New York, 2001).
12. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
13. Chen, H. *et al.* Limb and kidney defects in Lmx1b mutant mice suggest an involvement of LMX1B in human nail patella syndrome. *Nature Genet.* **19**, 51–55 (1998).

14. Poindexter, K., Nelson, N., DuBose, R. F., Black, R. A. & Cerretti, D. P. The identification of seven metalloproteinase-disintegrin (ADAM) genes from genomic libraries. *Gene* **237**, 61–70 (1999).
15. Mitelman, F., Mertens, F. & Johansson, B. A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genet.* **15**, 417–474 (1997).
16. The BAC Resource Consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**, 953–958 (2001).
17. Wooster, R. *et al.* Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12–13. *Science* **265**, 2088–2090 (1994).
18. Schuler, G. D. Sequence mapping by electronic PCR. *Genome Res.* **7**, 541–550 (1997).
19. Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* **28**, 352–355 (2000).
20. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
21. Struhl, K. Histone acetylation and transcriptional regulatory mechanisms. *Genes Dev.* **12**, 599–606 (1998).
22. Finnin, M. S. *et al.* Structures of a histone deacetylase homologue bound to the TSA and SAHA inhibitors. *Nature* **401**, 188–193 (1999).
23. Leippe, D. D. & Landsman, D. Histone deacetylases, acetoin utilization proteins and acetylpolymine amidohydrolases are members of an ancient protein superfamily. *Nucleic Acids Res.* **25**, 3693–3697 (1997).
24. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **28**, 263–266 (2000).
25. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28**, 231–234 (2000).

Acknowledgements

We thank G. Marth, S. Sherry, D. Landsman, D. Church and D. Lipman for suggestions and review of the manuscript.

Correspondence should be addressed to G.D.S. (e-mail: schuler@ncbi.nlm.nih.gov).

Mining the draft human genome

Ewan Birney*, Alex Bateman†, Michele E. Clamp† & Tim J. Hubbard†

* The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

† The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Now that the draft human genome sequence is available, everyone wants to be able to use it. However, we have perhaps become complacent about our ability to turn new genomes into lists of genes. The higher volume of data associated with a larger genome is accompanied by a much greater increase in complexity. We need to appreciate both the scale of the challenge of vertebrate genome analysis and the limitations of current gene prediction methods and understanding.

In this issue, accompanying the description of the sequence¹, there are nine data-mining papers that interrogate the genome from distinct biological perspectives. These range from broad topics—cancer², addiction³, gene expression⁴, immunology⁵ and evolutionary genomics⁶—to the more focused: membrane trafficking⁷, cytoskeleton⁸, cell cycle⁹ and circadian clock¹⁰. The findings reported by these authors are likely to be indicative of many people's experiences with the draft human genome: frustrating and rewarding in equal measures.

The current data set

The human genome—the first vertebrate genome sequence to be determined—seems likely to be quite representative of what we will find in other vertebrate genomes. It is around 30 times larger than the recently sequenced worm and fly genomes, and 250 times larger than that of yeast, the first eukaryotic genome to be sequenced¹¹. Despite its size, it seems likely to have only two or three times as many genes as the fly and worm genomes, with the coding regions of genes accounting for only 3% of the DNA. Repeat sequences form a large proportion of the remaining DNA, around 46%. These repeats may or may not have a function, but they are certainly characteristic of large vertebrate genomes. The rest of the sequence contains promoters, transcriptional regulatory sequences and other features, as yet unknown.

The International Human Genome Sequencing Consortium has been sequencing the genome in fragments of about 100–200 kilobases (kb). These fragments exist as bacterial artificial chromosome (BAC) clones, which are derived from sequences whose chromosomal location is known. Each newly generated sequence is deposited in the high-throughput genome sequence (HTGS) division of the International Nucleotide Database (GenBank/EMBL/DBJ) within 24 hours of being assembled and is assigned a unique identifier (its accession number). For the working draft, about 75% of clones are 'unfinished': each still consists of about 10–20 unassembled sequence fragments. Sequencing centres are continuously reading new sequence data from these clones until all the gaps are eliminated, at which point the sequence is declared 'finished'. As the HTGS entries are updated they retain the same accession numbers, but their version numbers increase.

There is a great deal of overlap between BAC clones, so it is typically more convenient to view a cleaned up version of the raw data, in which the sequences of the clones are correctly ordered and overlapped to remove redundancy and create a contiguous DNA sequence for each chromosome. These virtual chromosome sequences change continuously as gaps are closed and fragment ordering is refined.

Finding genes

With over 30 genomes sequenced, the casual observer could be forgiven for thinking that gene prediction, or annotation, was a problem filed neatly under 'solved'. Unfortunately this is far from

true. The large size of the genome makes finding the genes much more difficult. The protein-coding parts of human genes, called exons, are split into pieces in the genome and these pieces are separated by non-coding sequence called introns. Nearly all of the increase in gene size in human compared with fly or worm is due to the introns becoming much longer (about 50 kb versus 5 kb). The protein-encoding exons, on the other hand, are roughly the same size. This decrease in signal (exon) to noise (intron) ratio in the human genome leads to misprediction by computational gene-finding strategies.

Many methods for predicting genes are based on compositional signals that are found in the DNA sequence. These methods detect characteristics that are expected to be associated with genes, such as splice sites and coding regions, and then piece this information together to determine the complete or partial sequence of a gene. Unfortunately, these *ab initio* methods tend to produce false positives, leading to overestimates of gene numbers, which means that we cannot confidently use them for annotation. They also do not work well with unfinished sequence that has gaps and errors, which may give rise to frameshifts, when the reading frame of the gene is disrupted by the addition or removal of bases.

Thankfully, there is a wealth of data that we can use to produce more reliable gene predictions. Information on expressed sequences (expressed sequence tags (ESTs) and complementary DNAs) and proteins from humans and other organisms provide a more accurate resource for resolving gene structures against the vast genomic background. The most effective algorithms integrate gene-prediction methods with similarity comparisons. Such algorithms are integral to software programs such as GeneWise¹², Genomescan¹³ and Genie¹⁴, which provide accurate, automatic predictions, whereas BLAST or FASTA programs typically require considerable manual effort to determine the complete structure of a single gene.

The most powerful tool for finding genes may be other vertebrate genomes. Comparing conserved sequence regions between two closely related organisms will enable us to find genes and other important regions in both genomes with no previous knowledge of the gene content of either. The next couple of years should see the sequencing of the mouse, zebrafish and *Tetraodon* genomes. The preliminary sequence of *Tetraodon* has already proved useful in estimating gene numbers¹⁵, and shows much promise for the use of comparative genomics in gene prediction.

Resources available to the user

There are a number of resources currently available for perusing the human genome. 'Human Genome Central'¹⁶ attempts to gather together the most useful web sites (see <http://www.ensembl.org/genome/central/> or <http://www.ncbi.nlm.nih.gov/genome/central/>). The best starting point for the uninitiated will be a site such as those of NCBI, Ensembl or the University of Santa Cruz (UCSC). These sites offer a mixture of genomic viewers and web-searchable

datasets, and allow analysis of the human genome sequence without the need to run complex software locally.

For more involved analysis, it might be necessary to download some of the data locally. Useful downloadable sequence-oriented datasets include protein datasets (available from Ensembl and NCBI) and the assembled DNA sequence for regions of the genome, available at UCSC. Other genomic datasets are also available, such as the global physical map from The Genome Sequencing Center in St Louis and the single nucleotide polymorphism (SNP) database from NCBI. Raw sequence data is available from the International Nucleotide Database (GenBank/EMBL/DDBJ), but this data is generally more difficult to handle because it is very fragmentary, can contain contaminating non-human DNA and may include misleading information such as incorrect map assignment.

This loose network of sites will probably coalesce into a more coordinated network of sites offering informative web pages and resources. NCBI, Ensembl and UCSC are developing new, more accessible resources that will become available within the next year.

How to use the resources

There are two main ways to use the human genome sequence. First, we can look for a homologue of a protein that is known from another organism. For example, Clayton *et al.*² looked for relatives of the *Drosophila* period clock protein and found the three known relatives and a possible fourth cousin on chromosome 7. Or we can try and find all of the proteins belonging to a particular family—in ref. 4, Tupler *et al.* catalogue all homeobox domains⁴. The easiest way to approach these problems is to use a protein set. This sidesteps the frustration of predicting genes, but makes the researcher reliant on the quality of the predictions being provided. For most of the accompanying reports, a single protein set was the most useful resource provided. For example, Nestler *et al.* searched for G-protein receptor kinases³ using PSI-BLAST, which searches only protein datasets.

What are the potential pitfalls of the data? Human genes are hard to predict and are often fragmented. If each end of a query protein matches to a different predicted protein, we should suspect that the query sequence may in fact be two parts of a fragmented gene. The two matched human genes should be in the same or adjacent genomic locations. Pollard⁸ discovered that fragmentation complicated the analysis of myosin genes. In addition, the unfinished human genomic DNA may contain contamination, particularly from bacteria but also from other sources. Contaminating DNA is routinely removed from finished sequence, but some is still present in unfinished sequence. If the predicted gene matches a bacterial gene more closely than any vertebrate gene then it will almost always be a contaminant. Futreal *et al.*² were led up a blind alley for a week before they discovered that cDNA contamination in draft genomic

sequences was giving the false impression of multiple p53 proteins in the genome.

During the assembly of unfinished human genomic data it is possible to create artificial duplications, which can result in artefacts in the subsequent analysis. Very similar gene sequences found within the same clone may represent duplicate genes, but could also be the result of an assembly error. This also means that predicted protein sets may contain artificial duplications, leading to overestimation of the number of members in a family.

What does this analysis tell us? For Bock *et al.*⁷, the draft genome revealed a list of the molecular players involved in membrane trafficking, providing a platform for experiments that may complete our understanding of this area of biology. In contrast, Murray and Marks⁹ found no new cyclin-dependent kinases, indicating that they were all found by traditional experimental techniques. Futreal *et al.* had a similar experience for known cancer genes, but suggest that with new techniques the genome will provide new avenues of cancer research².

The interpretation of unfinished draft genomic data may seem like hard work. But it is something to become accustomed to, because we expect future vertebrate genomes to be released initially in draft form. The database providers must develop better ways of viewing the data; and researchers need to be educated in how to use them. That said, there are many undiscovered treasures in the current data set waiting to be found by intuition, hard work and experimental verification. Good luck, and happy hunting! □

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Futreal, A., Wooster, R., Kasprzyk, A., Birney, E. & Stratton, S. Cancer and genomics. *Nature* **409**, 850–852 (2001).
3. Nestler, E. J. & Landsman, E. Learning about addiction from the genome. *Nature* **409**, 834–835 (2001).
4. Tupler, R., Perini, G. & Green, M. R. Expressing the human genome. *Nature* **409**, 832–833 (2001).
5. Fahrner, A. M., Bazan, J. E., Papatianasiou, P., Nelms, K. A. & Goodnow, C. C. A genomic view of immunology. *Nature* **409**, 836–838 (2001).
6. Li, W.-H., Gu, Z., Wang, H. & Nekrutenko, A. Evolutionary analyses of the human genome. *Nature* **409**, 847–849 (2001).
7. Bock, J. B., Matern, H. T., Peden, A. A. & Scheller, R. H. A genomic perspective on membrane compartment organization. *Nature* **409**, 839–841 (2001).
8. Pollard, T. D. Genomics, the cytoskeleton and motility. *Nature* **409**, 842–843 (2001).
9. Murray, A. W. & Marks, D. Can sequencing shed light on cell cycling? *Nature* **409**, 844–846 (2001).
10. Clayton, J. D., Kyriacou, C. P. & Reppert, S. M. Keeping time with the human genome. *Nature* **409**, 829–831 (2001).
11. Goffeau, A. *et al.* The Yeast Genome Directory. *Nature* **387** (suppl.), 1–105 (1997).
12. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–548 (2000).
13. Burge *et al.* *Nature Genet.* (submitted).
14. Reese, M. G., Kulp, D., Tammana, H., Haussler, D. Genie—gene finding in *Drosophila melanogaster*. *Genome Res.* **10**, 529–538 (2000).
15. Crollius, H. R. *et al.* Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Res.* **10**, 939–949 (2000).
16. Genome website set up to help with sequence analysis. *Nature* **406**, 929 (2000).

Correspondence should be addressed to E.B. (e-mail: birney@ebi.ac.uk).

Keeping time with the human genome

Jonathan D. Clayton*, Charalambos P. Kyriacou* & Steven M. Reppert†

* Department of Genetics, University of Leicester, Leicester LE1 7RH, UK

† Laboratory of Developmental Chronobiology, MassGeneral Hospital for Children, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA

The cloning and characterization of 'clock gene' families has advanced our understanding of the molecular control of the mammalian circadian clock. We have analysed the human genome for additional relatives, and identified new candidate genes that may expand our knowledge of the molecular workings of the circadian clock. This knowledge could lead to the development of therapies for treating jet lag and sleep disorders, and add to our understanding of the genetic contribution of clock gene alterations to sleep and neuropsychiatric disorders. The human genome will also aid in the identification of output genes that ultimately control circadian behaviours.

Completion of the draft sequence of the human genome is an epic event, providing a treasure-trove of new genes for all aspects of biological investigation. For circadian clock research, 'snapshot' analysis of the draft sequence provides a glimpse of potential new members of canonical 'clock gene' families. Such candidate genes will add to the riches of the post-genomic era, ultimately leading to a wider understanding of the molecular control of the circadian clock.

The master circadian clock in mammals, in the suprachiasmatic nuclei (SCN) of the anterior hypothalamus, drives daily variations in many physiological and behavioural processes, such as the sleep–wake rhythm and daily variations in body temperature, hormone levels, cognition and memory¹. Dawn and dusk coordinate or entrain the circadian clock through neural pathways connecting the retina to the SCN, so that the master clock and its output rhythms do not drift from 24 h, but remain aligned with the solar day. Transient disruption of circadian timing following transmeridian flights leads to jet lag, and chronic alterations of the central clock mechanism in shift workers (around 25% of the working population) may contribute to poor health and sleep disorders. Finally, specific rhythm defects may be involved in neuropsychiatric illnesses.

The fruitfly *Drosophila melanogaster* has long been the animal model of choice for genetic analysis of circadian rhythms. In recent years, with the help of forward and reverse genetics, homology-based approaches and protein-interaction screens, homologues of most of the genes involved in the fly clock have been cloned in mammals² (Fig. 1, Table 1). Although there are compelling similarities between the general clock mechanisms of *Drosophila* and mice, gene duplication has led to reassignment of specific functions among structurally homologous components.

Analysis of the fly circadian clock appears to be close to genetic saturation, indicating that most of the cardinal genes may have been identified³. It is less clear whether the whole complement of clock genes has been identified in mammals, because expansion of gene families makes it likely that additional family members may await discovery. We searched the draft sequence of the human genome for new clock genes (see Supplementary Information for Methods).

A few words of caution are in order concerning the sequence analysis. Although we have identified new clock gene candidates, we have no information regarding their expression patterns or function. We detected all of the previously identified clock loci in mammals, confirming the near completeness of the human genome sequence data. This does not rule out the possibility, however, of further clock candidates surfacing from advanced searches of the finished genome sequence.

Genes involved in the core clock mechanism

The core clock mechanism of mammals is composed of interacting positive and negative transcriptional/translational feedback loops⁴ (Fig. 1). The pivotal components of this mechanism are two basic helix–loop–helix (bHLH)/PAS-containing transcription factors, CLOCK and BMAL1, which are assumed to pair through their protein-interactive PAS domains. PAS domains are found in a diverse family of proteins (PER, ARNT and SIM are the founder members) and are a common feature of transcription factors that are essential for the clock machinery of fungi, insects and mammals⁵.

CLOCK–BMAL1 heterodimers drive the rhythmic transcription of three *Period* genes (*mPer1–mPer3* in the mouse) and two

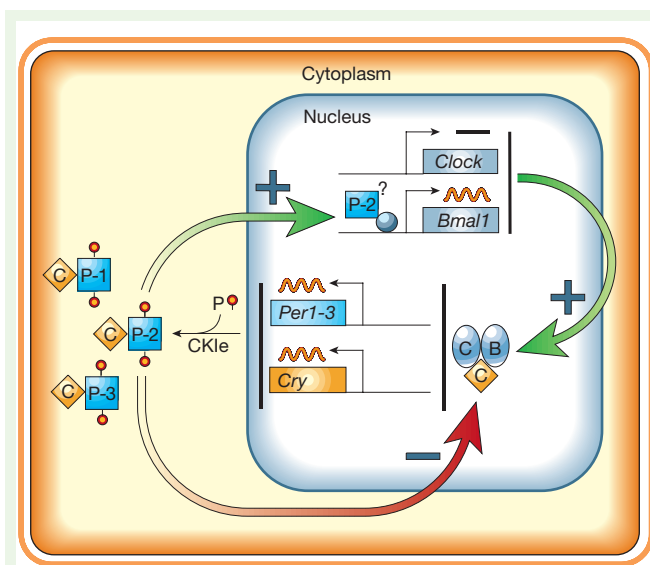


Figure 1 Model of circadian clockwork within the SCN. The clockwork is comprised of interacting positive (green) and negative (red) feedback loops. CLOCK (oval with C) and BMAL1 (oval with B) heterodimers activate (+) rhythmic transcription of *Cry* and *Per* genes. The CRY (C) and PER (P1–P3) proteins form complexes that are important for nuclear translocation of the PER proteins. The phosphorylation (P) of the PER proteins by CKIε may also regulate their cellular location and stability. PER2 may positively regulate transcription of *Bmal1* by acting as a co-activator. The nuclear-localized mCRY proteins interact directly with CLOCK and BMAL1 to negatively regulate (–) transcription mediated by CLOCK–BMAL1. PER1 may communicate light information to the molecular loops, as its RNA can be induced by light at dawn and dusk. PER3 may contribute to transducing the oscillation to output systems.

Cryptochrome genes (*mCry1* and *mCry2*). As the mPER and mCRY proteins are translated, they form PER–CRY complexes that are translocated to the nucleus. There the mCRY proteins act as negative regulators by directly interacting with CLOCK or BMAL1, or both, to inhibit transcription, forming a negative feedback loop. At the same time, mPER2 contributes to the transcription of *Bmal1*, which is rhythmically expressed with a peak phase opposite to that of *mPer/mCry*, forming a positive feedback loop. The push–pull action of the positive and negative feedback loops perpetuates the self-sustaining nature of the circadian clock.

Analysis of the draft sequence identified human *hPER1*, *hPER2*, *hPER3*, *hCRY1* and *hCRY2*. Moreover, an additional single-copy gene was identified on chromosome 7 with high sequence similarity to the *Per* family (Table 1). A fragment of genomic sequence is available, but no expressed sequence tag (EST) matches were obtained. Nevertheless, the putative translation of this fragment encodes 231 residues that encompass most of the PAS region and can be compared with those of other family members (see Supplementary Information Figs A and B). If '*hPER4*' can be confirmed as an expressed sequence rather than a pseudogene, the human genome project will have yielded a potentially exciting finding, because such a fourth *PER* will contribute to our understanding of the core clock mechanism. Even if it is not expressed, identification of an *hPER4* locus may help us to decipher the evolution of the *Per* gene family.

Our analysis identified *hBMAL1* (*hMOP3*) and *hCLOCK*, as well as two previously described members of this transcription factor family, *hMOP4* and *hMOP9* (also called *CLIF*)^{6–8}, which encode proteins closely related to hCLOCK and hBMAL1, respectively (Table 1; see Supplementary Information Figs C and D). We did not find any new genes related to these transcription factors. Because MOP4–BMAL1 and CLOCK–MOP9/CLIF heterodimers can activate transcription from E-boxes (CACGTG) *in vitro*, MOP4 and MOP9/CLIF have the potential to contribute to transcriptional regulation within the core clock mechanism. MOP9/CLIF is in fact expressed in the SCN^{7,8}.

Kinases important for clock function

Phosphorylation and proteolysis of clock proteins can determine their cellular location and stability, and are key ingredients for building time delays into the 24-h molecular mechanism³. In *Drosophila*, DOUBLETIME, which is closely related to mammalian casein kinase I epsilon (CKIε), phosphorylates PER, thereby influencing PER turnover³. Studies of the *tau* mutation in Syrian hamsters (a spontaneous, semidominant mutation leading to marked shortening of the circadian period) reveal that it encodes a missense mutation within CKIε that renders the mutant enzyme deficient in its ability to phosphorylate the mPER proteins⁹.

Casein kinase I delta (CKIδ) is highly homologous to CKIε (76% identical at the amino-acid level) and efficiently binds and phosphorylates mPER1 *in vitro*¹⁰. Our database search revealed another human casein kinase gene, '*hCKIε2*'. This gene is detected in three partially overlapping ESTs that could encode a carboxy-terminal fragment highly homologous to hCKIε (91% identical over 73 residues, with two of the substitutions in or near putative autophosphorylation sites; see Supplementary Information Figs E and F). Three other novel *hCK1* genes were found, one similar to chicken CKIγ1 (which we name *hCKIγ1*), and two related to *hCKIα1*, which we have called *hCKIα2* and *hCKIα3* (see Supplementary Information Figs E and F). Functional analysis of CKIε2 will certainly add to the complexity inherent in the phosphorylation events that are important for the clockwork.

The timeless mystery

The *timeless* (*tim*) gene is essential for circadian function in *Drosophila*³. Putative homologues of *Drosophila tim* have been identified in both mice and humans (*mTim* and *hTIM*, respectively),

but placing the homologue within the murine molecular mechanism has proved difficult². Database searches of the completed *Drosophila* genome show that mammalian TIM is not the true orthologue of *Drosophila* TIM, but is the likely orthologue of a newly described fly gene, *timeout* (also called *tim-2*)^{11,12}. As the core clock functions of TIM in the fly have been usurped by other clock-relevant genes in the mouse², there is no obvious requirement for a mammalian equivalent. Our analysis of the human genome sequence did not reveal a second *Timeless* homologue, consistent with this idea.

Analysis of the completed genome of the worm *Caenorhabditis elegans* provides additional insight into the *timeless* mystery. These animals are behavioural outliers, because no circadian rhythms have been reported in the worm. We might reasonably predict that some of the canonical clock genes would be absent from *C. elegans*. In fact, our searches of the worm databases did not detect any genes with similarity to cryptochromes. Significant hits were obtained for both *clock* and *bmal1* to the same *C. elegans* gene *aha-1* (an *hARNT* homologue; Table 1), with the similarity far greater for *bmal1*. The *C. elegans* gene *lin-42*, which is required for postembryonic development, shows some similarity in its PAS region to the *Drosophila* and mammalian PERs¹³.

Table 1 Clock genes in metazoans

Fruitfly (<i>Drosophila melanogaster</i>)	Worm (<i>Caenorhabditis elegans</i>)	Mouse (<i>Mus musculus</i>)	Human (<i>Homo sapiens</i>)
Period			
–	–	<i>mPer1</i> [11B] SWALL: O35973	<i>hPER1</i> [17p13.1] SWALL: O15534
<i>per</i> SWALL: P07663	<i>lin-42</i> SWALL: P91313	<i>mPer2</i> SWALL: O54943	<i>hPER2</i> [2] SWALL: O15055
–	–	<i>mPer3</i> SWALL: O70361	<i>hPER3</i> [1] SWALL: Q9UGU8
–	–	–	<i>hPER4</i> [7] gb: AC027390.3
Timeless			
<i>tim</i> SWALL: P49021	–	–	–
<i>timeout/tim2</i> SWALL: Q9NG27	<i>tim-1</i> SWALL: Q9xw65	<i>mTim*</i> [10:18.0] SWALL: Q9Z0E7	<i>hTIM</i> [12q12] SWALL: O94802
Clock/Bmal1			
<i>Clock/Jerk</i> SWALL: O61735	<i>aha-1</i> SWALL: Q18141	<i>mClk</i> [5:43.0] SWALL: O08785 <i>mNPAS2</i> [1.20] SWALL: P97460	<i>hCLK</i> [4q12] SWALL: O15516 <i>hMOP4</i> [2p11.2] SWALL: Q99742
<i>cycle</i> SWALL: O61734	<i>aha-1</i> SWALL: Q18141	<i>mBmal1</i> (<i>mMOP3</i>) [7.52] SWALL: Q9WTL8 SWALL: O88295	<i>hBMAL1</i> (<i>hMOP3</i>) [11p15] SWALL: O00327
Casein Kinase 1ε			
<i>doubletime*</i> SWALL: O76324	<i>CK1δ</i> SWALL: Q20471	<i>tau</i> (<i>mCK1ε</i>) [15:E1] SWALL: Q9JMK2	<i>hCK1ε</i> [22q13.1] SWALL: P49674 <i>hCK1ε2</i> aw673521 aa934662 aa001679
Cryptochrome			
<i>cry</i> SWALL: O77059	–	<i>mCry1</i> [10C] SWALL: P97784	<i>hCRY1</i> [12q23] SWALL: Q16526
–	–	<i>mCry2</i> [2E] SWALL: Q9R194	<i>hCRY2</i> [11] SWALL: O75148

The results are from searches of the draft human genome database with members of the canonical clock genes from fly and mouse. Genes on the same row are the most similar between species: for example, *dper* is most similar in sequence to *mPer2*, and worm *CK1δ* is the closest to fly and mammalian *CKIε*. Dashes indicate that there is no similar sequence in the corresponding database. No similarities to either *hPER4* or *hCKIε2* were found in mouse genomic or expressed sequence tag databases. For the human genome, chromosome positions (if known) are given in square brackets, and novel, closely related family members discovered in this search are given in bold. Underlined genes, when mutated, give clock phenotypes. SWALL, SwissProt accession numbers; gb, GenBank accession numbers. See Supplementary Information for detailed Methods.

* Loci with lethal mutations associated.

Most notably, the worm has only one *timeless* gene, *tim-1*, which is highly homologous to *Drosophila timeout/tim-2* and mammalian *Tim*, but substantially removed from the clock-relevant *timeless* in the fly¹³. This might suggest that the closely related worm *tim-1*, mouse/human *Tim* and fly *timeout/tim-2* genes are descendants of an ancestral *timeless* gene that duplicated in the arthropod lineage after the split with nematodes and vertebrates¹². The clock-relevant *timeless* duplication could then have rapidly diverged to take on a dedicated role in the circadian machinery of insects.

Impact of the human genome on chronobiology

This snapshot analysis of the draft sequence of the human genome provides candidate sequences that could increase the number of clock genes in mammals. Aided by further genomic and post-genomic analyses, this could provide new opportunities for pharmacological manipulation of the human clockwork with 'chronotherapies' targeted at improving the treatments for those who work shifts or suffer from jet lag, and those with clock-related sleep or psychiatric problems.

Recently, an autosomal dominant sleep disorder characterized by debilitating early sleep onset and offset was shown to involve a missense mutation in the CKIε binding region of *hPER2*; the mutation leads to decreased phosphorylation of *hPER2* by CKIε *in vitro*¹⁴. Linkage analysis identified a candidate locus in the affected family, and the chromosomal location of known clock genes from the human genome provided a short-cut in the search. Identified polymorphisms within human clock genes may aid genetic analysis of other rhythm-related traits, such as the early riser 'lark' and the late riser 'owl' phenotypes¹⁵.

The completed human genome sequence will also have a significant impact on the study of circadian output systems. One of the most profound discoveries of the mammalian clock gene 'era' has been the finding that clock genes in mice and humans are widely expressed throughout the body, where they seem to participate in local circadian oscillator function². These peripheral oscillators are dampened, and require cyclic input from the master clock in the SCN for sustained rhythmicity. Local clock-controlled genes in peripheral organs would provide the flexibility needed to control the circadian rhythms present in these tissues.

Using DNA chip technology, expression profiling of rhythmic gene expression in the SCN and in peripheral tissues will identify candidate output genes and their temporal contribution to local physiology. The completed human genome sequence will then play its part in the identification of these output genes and accelerate fulfilment of one of the great promises of circadian rhythm research: providing a complete understanding of the cellular and molecular events connecting clock genes to circadian behaviour.

1. Klein, D. C., Moore, R. Y. & Reppert, S. M. (eds) *Suprachiasmatic Nucleus: The Mind's Clock* (Oxford Univ. Press, New York, 1991).
2. Reppert, S. M. & Weaver, D. R. Comparing clockworks: mouse versus fly. *J. Biol. Rhythms* **15**, 357–364 (2000).
3. Young, M. W. Circadian rhythms. Marking time for a kingdom. *Science* **288**, 451–453 (2000).
4. Shearman, L. P. *et al.* Interacting molecular loops in the mammalian circadian clock. *Science* **288**, 1013–1019 (2000).
5. Dunlap, J. C. Molecular bases for circadian clocks. *Cell* **96**, 271–290 (1999).
6. Hogenesch, J. B., Gu, Y. Z., Jain, S. & Bradfield, C. A. The basic-helix-loop-helix-PAS orphan MOP3 forms transcriptionally active complexes with circadian and hypoxia factors. *Proc. Natl Acad. Sci. USA* **95**, 5474–5479 (1998).
7. Hogenesch, J. B. *et al.* The basic helix-loop-helix-PAS protein MOP9 is a brain-specific heterodimeric partner of circadian and hypoxia factors. *J. Neurosci.* **20**, RC83 (2000).
8. Maemura, K. *et al.* CLIF, a novel cycle like factor, regulates the circadian oscillation of plasminogen activator inhibitor-1 gene expression. *J. Biol. Chem.* **275**, 36847–36851 (2000).
9. Lowrey, P. L. *et al.* Positional syntenic cloning and functional characterization of the mammalian circadian mutation *tau*. *Science* **288**, 483–491 (2000).
10. Vielhaber, E., Eide, E., Rivers, A., Gao, Z. H. & Virshup, D. M. Nuclear entry of the circadian regulator mPER1 is controlled by mammalian casein kinase 1ε. *Mol. Cell. Biol.* **10**, 4888–4899 (2000).
11. Gotter, A. L. *et al.* A time-less function for mouse *Timeless*. *Nature Neurosci.* **3**, 755–756 (2000).
12. Benna, C. *et al.* A second *timeless* gene in *Drosophila* shares greater sequence similarity with mammalian *tim*. *Curr. Biol.* **10**, R513 (2000).
13. Jeon, M., Gardner, H. F., Miller, E. A., Deshler, J. & Rougvie, A. E. Similarity of the *C. elegans* developmental timing protein LIN-42 to circadian rhythm proteins. *Science* **286**, 1141–1146 (1999).
14. Toh, K. L. *et al.* An *hPer2* phosphorylation site mutation in familial advanced sleep-phase syndrome. *Science*, 11 January 2001 (10.1126/science.1057499).
15. Katzenberg, D. *et al.* A CLOCK polymorphism associated with human diurnal preference. *Sleep* **21**, 569–576 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank A. Kasprzyk at the European Bioinformatics Institute for assistance. This work was funded by an International Human Frontiers Grant.

Correspondence and requests for materials should be addressed to S.M.R. (e-mail: reppert@helix.mgh.harvard.edu).

Expressing the human genome

Rossella Tupler*†, Giovanni Perini‡ & Michael R. Green*

* Howard Hughes Medical Institute, Programs in Gene Function and Expression and Molecular Medicine, University of Massachusetts Medical School, 373 Plantation St, Worcester, Massachusetts 01605, USA

† Biologia Generale e Genetica Medica, Università degli Studi di Pavia, via Forlanini 14, 27100 Pavia, Italy

‡ University of Bologna, Dipartimento di Biologia Evoluzionistica e Sperimentale, 40126, Via Selmi 3, Bologna, Italy

We have searched the human genome for genes encoding new proteins that may be involved in three nuclear gene expression processes: transcription, pre-messenger RNA splicing and polyadenylation. A plethora of potential new factors are implicated by sequence in nuclear gene expression, revealing a substantial but selective increase in complexity compared with *Drosophila melanogaster* and *Caenorhabditis elegans*. Although the raw genomic information has limitations, its availability offers new experimental approaches for studying gene expression.

The availability of the human and other genome sequences will revolutionize all fields of biomedical research. But, as the genome is itself the object of gene expression, the impact may be particularly profound for those of us studying this process. Here we preview what the human genome has to offer to those interested in gene expression.

Gene expression comprises various events from the initial transcription of a gene in the nucleus to the translation of mRNA in the cytoplasm. Here we focus on three steps that occur in the nucleus: transcription, pre-mRNA splicing and 3' end formation. All three involve the recognition of a nucleic acid (DNA or RNA) that serves as a scaffold for a multiprotein complex in which the relevant reaction (transcription, splicing or 3' end formation) occurs. Research in each is aimed at identifying all the relevant components and elucidating how the reaction is controlled.

General transcription factors

Factors involved in the transcription of eukaryotic protein-coding genes by RNA polymerase II fall into two groups: general (or basic) transcription factors (GTFs) and transcriptional activators. GTFs are required for accurate transcription initiation *in vitro*¹. They include RNA polymerase II itself and at least six GTFs: TFIID, TFIIA, TFIIB, TFIIE, TFIIF and TFIIH. The GTFs assemble on the promoter to form a preinitiation complex (PIC). Of the GTFs, TFIID is the primary sequence-specific DNA-binding component; it initiates PIC assembly by interaction with the TATA box. TFIID is composed of the TATA-box-binding protein (TBP) and multiple TBP-associated factors (TAFs)². This basic transcription machinery has, in general, been highly conserved from yeast to human.

We have searched the human genome sequence for GTFs. Consistent with the *Drosophila*, *C. elegans* and *Saccharomyces cerevisiae* genomes³, the human genome contains single-copy genes encoding the components of RNA polymerase II, TFIIB, TFIIE, TFIIF and TFIIH, in general without evidence for related genes (see Supplementary Information Table 1). A potentially important exception is the presence of three genes related to *cdk7*, a cyclin-dependent kinase that is associated with TFIIH⁴.

We found gene sequences related to many of the TFIID subunits, including TBP, TFIIA and several TAFs, indicating that the potential diversity of human TFIID is much greater than that of *Drosophila*³ (see Supplementary Information Table 1). For example, we identified six human genes related to TAF32, but no *Drosophila* genes related to the homologous TAF40. It was thought that all metazoans possess a single gene for TBP, with a second gene encoding a TBP-like factor (TLF). Unusually, *Drosophila* contains a third TBP-related gene (TRF1)⁵. However, our searches reveal that humans also have a third TBP-related sequence on chromosome 14 (see Supplementary Information Fig. 1).

Transcriptional activators

Transcriptional activity is strongly stimulated by promoter-specific activators. In general, these are sequence-specific DNA-binding proteins whose recognition sites are present in target promoters. Activators have been classified into families on the basis of their DNA-binding domains. A search of the human genome sequence revealed more than 2,000 hypothetical genes that encode transcriptional activators (Fig. 1). The C2H2 zinc finger proteins form the largest family (around 900 members), and this is also the largest family of activators in *Drosophila*, *C. elegans* and *S. cerevisiae*. There are around twice as many basic region leucine zipper (bZIP) proteins, nuclear receptors and helix-loop-helix proteins in the human genome as in the *Drosophila* genome, and 5–10 times more than in the *C. elegans* and *S. cerevisiae* genomes (Fig. 1).

We performed more extensive searches on the bZIP protein superfamily of transcriptional activators. We have aligned these new bZIP proteins on the basis of sequence relatedness to existing members of the bZIP subfamilies, such as Jun, Fos, ATF, CREB and c/EBP (see Supplementary Information Fig. 2). This revealed 18 new bZIP genes, primarily belonging to the Fos and CREB families. Three new genes belong to the small TEF/HLF family, which shares strong similarity with the *C. elegans* gene *ces-2*, which is involved in the control of programmed cell death^{6,7}. This search illustrates how the human genome sequence will provide many new factors that may be involved in gene expression, although their roles remain unknown.

Pre-mRNA splicing

Pre-mRNA splicing occurs in a large, dynamic complex, the spliceosome. To assemble the spliceosome, four small nuclear ribonucleoprotein (snRNP) particles (U1, U2, U5 and U4/U6) and many non-snRNP proteins interact with pre-mRNA in an ordered pathway^{8,9}. In metazoans, spliceosome assembly begins with recognition of the 5' splice site by U1 snRNP, and of the polypyrimidine (Py)-tract by the U2 snRNP auxiliary factor, U2AF¹⁰.

We have searched the human genome sequence for several protein-splicing factors that act early during spliceosome assembly. The results reveal significantly greater complexity than is found in *Drosophila*¹¹ (see Supplementary Information Table 2). Of particular interest are several new genes closely related to the small subunit of U2AF, U2AF³⁵ (see Supplementary Information Fig. 3).

Polyadenylation

Eukaryotic mRNAs have a 3' poly(A) tail, around 200 nucleotides long, which is added post-transcriptionally following endonucleolytic cleavage of the pre-mRNA. Addition of poly(A) is directed by a polyadenylation sequence, AAUAAA, located 10–30 bases upstream of the polyadenylation site. Polyadenylation requires

various protein components, including a cleavage/polyadenylation specificity factor (CPSF), a cleavage stimulatory factor (CstF), other cleavage factors and a poly(A) polymerase (PAP)¹². Several poly(A)-binding proteins (PABs) bind to the mature poly(A) tail.

We identified genes related to factors involved in mRNA 3' end formation, including PAP and PAB (see Supplementary Information Table 3). It was known that there were multiple, alternatively spliced variants of PAP, but the existence of several PAP-related gene sequences was unexpected and increases the potential diversity of this enzyme. These results suggest that the polyadenylation machinery of humans is more complex than that of *Drosophila*.

Limitations of the genomic information

Although these searches highlight the power of the new genomic information, they also reveal important limitations. In particular, the existence of a related gene sequence does not mean that there is a corresponding protein: the sequence could be a non-expressed pseudogene. Indeed, some new gene sequences contain stop codons or lack introns. We know that some of the related gene sequences are expressed, however, as they are present in expressed sequence tag (EST) databases. Even if a related sequence is an expressed gene, we do not know whether the two related genes are simultaneously expressed in the same cell, or are differentially expressed—for example, in a tissue- or development-specific manner. Expression studies will be required to complement genomic information. A final caveat is that many of the factors are components of multi-subunit complexes. Sometimes the same factor is present in multiple complexes whose activities differ substantially. Thus, the full value of the genomic information can be realized only when it is coupled with appropriate biochemical studies.

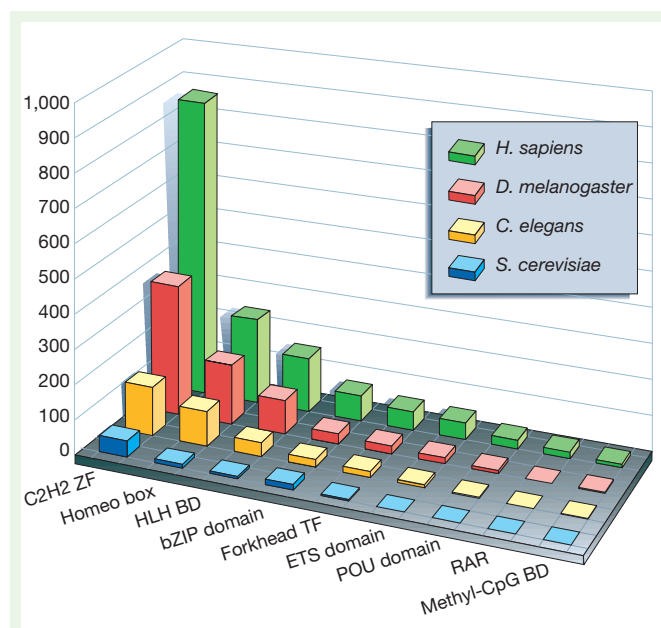


Figure 1 Genome-wide comparison of transcriptional activator families in eukaryotes. The relative sizes of transcriptional activator families among *Homo sapiens*, *D. melanogaster*, *C. elegans* and *S. cerevisiae* are indicated, derived from an analysis of eukaryotic proteomes using the INTERPRO database, which incorporates Pfam, PRINTS and Prosite. The transcription factors families shown are the largest of their category out of the 1,502 human protein families listed by the IPI.

New approaches for studying gene expression

Full eukaryotic genome sequences will allow new experimental strategies to study gene expression. In the 'classical' pre-genomic strategy, factors involved in gene expression were identified through biochemical or genetic experiments that focused on a specific gene expression process (for example, a transcription pathway). The human genome sequence contains many new factors whose sequences suggest roles in gene expression but whose precise activities and functions are unknown. Therefore, a 'post-genomic' approach can start with the new gene (or its protein product) and determine its activity and the pathway in which it participates.

Consider the new bZIP proteins. Clues to the biological processes in which they participate may be obtained by determining their tissue expression patterns, elucidating their DNA binding specificities¹³ or mapping their chromosomal sites of occupancy¹⁴. Ultimately, understanding the function of a transcriptional activator will require identification of the genes that it controls. This requires derivation of appropriate cell lines or animal models in which the protein can be expressed (or inactivated) in a regulated fashion. The consequences of its expression (or inactivation) on the transcription of specific genes is then monitored. In the most powerful version of this approach, target gene transcription can be assayed in parallel using high-density DNA microarrays¹⁵.

By performing focused searches of the draft human genome sequence, we have identified many new factors that may be involved in transcription, splicing and mRNA 3' end formation. The increased complexity of the human gene expression machinery implies that gene expression may be particularly important in shaping human development and physiology. □

- Orphanides, G., Lagrange, T. & Reinberg, D. The general transcription factors of RNA polymerase II. *Genes Dev.* **10**, 2657–2683 (1996).
- Burley, S. K. & Roeder, R. G. Biochemistry and structural biology of transcription factor IID (TFIID). *Annu. Rev. Biochem.* **65**, 769–799 (1996).
- Aoyagi, N. & Wassarman, D. A. Genes encoding *Drosophila melanogaster* RNA polymerase II general transcription factors: diversity in TFIIA and TFIID components contributes to gene-specific transcriptional regulation. *J. Cell. Biol.* **150**, F45–F49 (2000).
- Coin, F. & Egly, J. M. Ten years of TFIID. *Cold Spring Harb. Symp. Quant. Biol.* **63**, 105–110 (1998).
- Berk, A. J. TBP-like factors come into focus. *Cell* **103**, 5–8 (2000).
- Inaba, T. *et al.* Reversal of apoptosis by the leukaemia-associated E2A-HLF chimaeric transcription factor. *Nature* **382**, 541–544 (1996).
- Metzstein, M. M., Hengartner, M. O., Tsung, N., Ellis, R. E. & Horvitz, H. R. Transcriptional regulator of programmed cell death encoded by *Caenorhabditis elegans* gene *ces-2*. *Nature* **382**, 545–547 (1996).
- Burge, C. B., Tuschl, T. H. & Sharp, P. A. in *The RNA World* 2nd edn (eds Gesteland, R. F., Cech, T. R. & Atkins, J. R.) 525–560 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1999).
- Staley, J. P. & Guthrie, C. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92**, 315–326 (1998).
- Reed, R. Initial splice-site recognition and pairing during pre-mRNA splicing. *Curr. Opin. Genet. Dev.* **6**, 215–220 (1996).
- Mount, S. M. & Salz, H. K. Pre-messenger RNA processing factor in the *Drosophila* genome. *J. Cell. Biol.* **150**, F37–F43 (2000).
- Colgan, D. F. & Manley, J. L. Mechanism and regulation of mRNA polyadenylation. *Genes Dev.* **11**, 2755–2766 (1997).
- Ouellette, M. M. & Wright, W. E. Use of reiterative selection for defining protein-nucleic acid interactions. *Curr. Opin. Biotechnol.* **6**, 65–72 (1995).
- Orlando, V. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem. Sci.* **25**, 99–104 (2000).
- Young, R. A. Biomedical discovery with DNA arrays. *Cell* **102**, 9–5 (2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

M.R.G. is an investigator of the Howard Hughes Medical Institute. This work was supported in part by a grant from NIH to M.R.G.

Correspondence and requests for materials should be addressed to M.R.G. (e-mail: michael.green@umassmed.edu).

Learning about addiction from the genome

Eric J. Nestler* & David Landsman†

* Department of Psychiatry, The University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9070, USA

† National Center for Biotechnology Information, National Library of Medicine, Computational Biology Branch, Building 45, Room 6AN12J, 45 Center Drive, MSC 6510, Bethesda, Maryland 20892-6510, USA

Drug addiction can be defined as the compulsive seeking and taking of a drug despite adverse consequences. Although addiction involves many psychological and social factors, it also represents a biological process: the effects of repeated drug exposure on a vulnerable brain. The sequencing of the human and other mammalian genomes will help us to understand the biology of addiction by enabling us to identify both genes that contribute to individual risk for addiction and those through which drugs cause addiction. We illustrate this potential impact by searching a draft sequence of the human genome for genes related to desensitization of receptors that mediate the actions of drugs of abuse on the nervous system.

To understand addiction, it is important to define the types of molecular and cellular adaptation at the levels of neurons and synapses that account for tolerance, sensitization and dependence, which are often used to define an addicted state. Tolerance describes diminishing sensitivity to a drug's effects after repeated exposure; sensitization describes the opposite. Dependence is an altered physiological state caused by repeated drug exposure, which leads to withdrawal when drug use is discontinued. Each is seen in human addicts and is believed to contribute to continued drug use during addiction^{1,2}. Considerable progress has been made in identifying the molecular and cellular adaptations that mediate these processes³.

Challenges in addiction

A cardinal feature of addiction is its chronicity. Individuals can experience intense craving for drug and remain at increased risk for relapse even after years of abstinence, so addiction must involve very stable changes in the brain. But it has been difficult to identify such changes at the molecular, cellular or circuit levels. The molecular and cellular adaptations related to tolerance, sensitization and dependence do not persist long enough to account for the more stable behavioural changes associated with addiction.

This challenge is analogous to that faced in the field of learning and memory where there has been increasing appreciation for the role of learning-related processes in addiction^{1,2}. Many molecular and cellular models of learning have been discovered and some have been related to simple forms of learning behaviour^{4,5}. But little information is available concerning the molecular and cellular basis of essentially life-long memories. Proposed changes in synaptic structure, or in chromatin organization, remain speculative.

Another challenge is to identify the variations in specific genes that make some individuals vulnerable to addiction and others relatively resistant⁶. Epidemiological studies indicate that 40–60% of an individual's risk for an addiction, whether it is to alcohol, opiates or cocaine, is genetic. This is consistent with the widely differing sensitivity to drugs of abuse, including preferences to self-administer drug, among inbred rodent strains and lines⁷. However, we have not identified the specific genes involved in humans or animal models; nor do we understand with any specificity how external factors (including stress or drugs themselves) interact with those genetic variations to produce addiction.

Drugs of abuse seem to cause addiction by acting on evolutionarily old brain circuits. These circuits, which comprise several areas of the limbic system (for example, nucleus accumbens, amygdala and prefrontal cortex), regulate an organism's responses to natural reinforcers, such as food, drink, sex and social interaction^{1,2}. The loss of control that addicts show with respect to drug seeking and taking may relate to the ability of drugs of abuse

to commandeer these natural reward circuits and disrupt an individual's motivation and drive for normal reinforcers. There is evidence that 'natural addictions', such as overeating, pathological gambling, compulsive shopping and perhaps excessive exercise, may involve analogous mechanisms. A major focus of current research is to explore the neurobiology of these conditions and the influence of genetic factors in their development.

Impact of sequencing the human genome

We now know the initial targets for most drugs of abuse, as well as some of the molecular and cellular adaptations that occur in limbic brain circuits in response to repeated exposure. The draft sequence of the human genome indicates the diversity of the molecular components that have been implicated in addiction. For example, cocaine acts on the re-uptake transporters for dopamine and other monoamine neurotransmitters; we will soon know how many subtypes of such transporters are expressed in humans.

Another example is provided by genes whose products regulate desensitization of G-protein-coupled receptors. Such receptors are the initial targets for several drugs of abuse: opiates are agonists at opioid receptors, cannabinoids are agonists at cannabinoid receptors, and hallucinogens are partial agonists at serotonin 5HT_{2A} receptors. Dopamine receptors, which are indirectly activated by cocaine and other stimulants through potentiation of dopaminergic transmission, are also G-protein-coupled. The sensitivity of G-protein-coupled receptor signalling is controlled by complex regulatory processes (Fig. 1). Given the importance of changes in receptor sensitivity in addiction, it is not surprising that mechanisms governing receptor sensitivity have been implicated in regulating responses to drugs of abuse and in models of addiction^{3,8–10}.

A critical step in exploring such mechanisms is to identify all of the potential gene products that could be involved. Table 1 shows the results of an analysis of the current human protein dataset for some of the genes that regulate receptor desensitization: G-protein-receptor kinases (GRKs), arrestins, phosphatases and regulators of G-protein signalling (RGS proteins). GRKs phosphorylate ligand-bound receptors, enabling association of the receptors with arrestins^{8,9}. This seems functionally to uncouple receptors from their G proteins, perhaps through receptor internalization. Phosphatases also appear to alter receptor/G-protein interactions by regulating the availability of G-protein $\beta\gamma$ -subunits¹¹. RGS proteins serve as GTPase activating proteins for G-protein α -subunits and thereby alter the kinetics of a receptor-mediated response¹².

Our analysis of these gene families reveals new candidate members for each, and in several cases many, which can be investigated for their roles in addiction. For example, the specificity of the GRK–arrestin system for various types of G-protein-coupled receptor is not yet known. Knowledge of the full complement of GRKs and

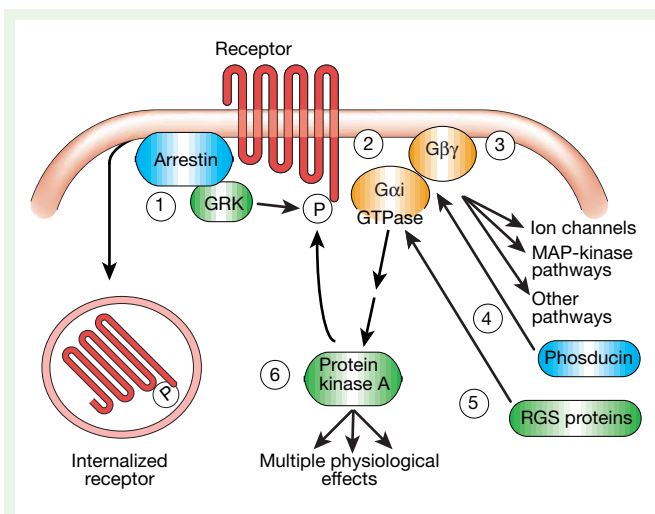


Figure 1 Possible mechanisms of drug-induced changes in the sensitivity of a G_i -coupled receptor (for example, opioid, cannabinoid and certain dopamine receptors). Drug-induced adaptations in the efficacy of receptor– G_i coupling could contribute to drug tolerance or sensitization. A possible mechanism is altered phosphorylation of the receptor by GRKs or its subsequent association with arrestins (1). Other possibilities include alterations in G-protein α - (2) or $\beta\gamma$ -subunits (3) or in other proteins (for example, phosphoducins (4) or RGS proteins (5)) that modulate G protein function. Phosphorylation of the receptor by protein kinase A (6) or other kinases represents another potential mechanism. Also shown is agonist-induced receptor internalization, which may be mediated by receptor phosphorylation. From ref. 3.

arrestins in humans allows us to evaluate the role of each GRK and arrestin subtype in regulating the sensitivity of the opioid, cannabinoid, serotonergic and dopaminergic receptors implicated in addiction. Similarly, the identification of novel subtypes of phosphoducins and RGS proteins makes it possible to determine which are expressed in neurons that mediate responses to drugs of abuse and which are involved in longer-term adaptations to drug exposure. Analogous efforts aimed at the receptors and G-protein subunits, as well as the ion channels (for example, inwardly rectifying potassium channels) that are regulated by the G proteins³, will provide more complete understanding of how drugs of abuse alter receptor signalling to produce tolerance and sensitization.

Access to the complete human genome sequence will also help efforts to identify addiction vulnerability genes. One of the major obstacles to such efforts has been the technical difficulty of moving from genetic linkage analyses to identification of specific genes⁶. The mouse genome sequence will similarly improve the efficiency of identifying addiction vulnerability genes in quantitative trait locus analyses of animal models⁷.

The power of genomics

Genomics (and the related proteomics) will provide powerful tools to identify the genes and gene products that are altered by repeated exposure to drugs of abuse and by external factors (for example, stress and drug-associated environmental stimuli) that influence the development of addiction. For example, DNA array technology makes it feasible to investigate thousands of gene products simultaneously after drug exposure. By combining genomic and proteomic tools with increasingly sophisticated models of addiction in animals, it will be possible to identify patterns of altered gene expression that are associated with particular features of the addicted state, such as tolerance, sensitization, dependence, craving and relapse.

Detailed analysis of the genome will also reveal how genes are organized and transcribed, and indicate the regulatory elements that control their expression. In addition, such analysis will show the exonic and intronic sequences of individual genes and teach us how to predict their processing into splice variants. We will similarly

Table 1 Diversity in genes that regulate desensitization of G-protein-coupled receptors

Gene	PSIBLAST ¹³ hits	SSEARCH ^{14,15} hits	New sequences PSIBLAST	New sequences SSEARCH	EST match
G protein-receptor kinases	576	361	91	57	48
Arrestins	5	5	1	1	1
RGS proteins	31	23	9	4	5
Phosphoducins	14	9	8	6	4

We queried the human protein dataset (42,227 sequences from the ENSEMBL dataset, <http://www.ensembl.org/>) with four human protein sequences: β -adrenergic receptor kinase-1 (SP:P25098; also called G protein-receptor kinase-2), β -arrestin (SP:P49407), regulator of G-protein signalling-5 (SP:O15539) and phosphoducin (gi:4505653). The columns labelled PSIBLAST and SSEARCH hits are the number of statistically significant hits obtained with each search method. To verify the expression of candidate gene products, each predicted human protein sequence in the 'New sequences PSIBLAST' column was compared to the expressed sequence tag (EST) database using TBLASTN with default parameters. The total number of protein sequences that showed perfect matches to at least one human EST is shown in the column labelled 'EST match'. (Some proteins gave perfect matches to ESTs in mouse or rat but not to human.) For details see Supplementary Information.

better appreciate how individual translated proteins are processed into multiple polypeptide products. Knowledge of such rules will greatly facilitate research into addiction. For example, as we elucidate the rules of gene transcription, we may be able to compile a list of genes that are potentially regulated by a transcription factor implicated in addiction, on the basis of the presence of the appropriate response element within their regulatory regions.

Finally, animal models of addiction are well developed, in contrast to other psychiatric abnormalities (for example, depression, bipolar disorder and schizophrenia) for which animal models are less straightforward. As a result, genomic studies of addiction might lead the way in identifying the molecular and cellular basis of complex behavioural states. A better understanding of the biology of addiction should help us to understand the mechanisms underlying symptoms of depression, anxiety and other disorders that overlap with those of addiction. Such understanding might also lead to appreciation of the factors that regulate normal variations in motivation, reward and mood. □

1. Koob, G. F., Sanna, P. P. & Bloom, F. E. Neuroscience of addiction. *Neuron* **21**, 467–476 (1998).
2. Wise, R. A. Drug-activation of brain reward pathways. *Drug Alcohol Depend.* **51**, 13–22 (1998).
3. Nestler, E. J. & Aghajanian, G. K. Molecular and cellular basis of addiction. *Science* **278**, 58–63 (1997).
4. Kandel, E. R. Genes, synapses, and long-term memory. *J. Cell. Physiol.* **173**, 124–125 (1997).
5. Malenka, R. C. & Nicoll, R. A. Long-term potentiation—a decade of progress? *Science* **285**, 1870–1874 (1999).
6. Nestler, E. J. Genes and addiction. *Nature Genet.* **26**, 277–281 (2000).
7. Crabbe, J. C., Phillips, T. J., Buck, K. J., Cunningham, C. L. & Belknap, J. K. Identifying genes for alcohol and drug sensitivity: recent progress and future directions. *Trends Neurosci.* **22**, 173–179 (1999).
8. Zhang, J. *et al.* Role for G protein-coupled receptor kinase in agonist-specific regulation of mu-opioid receptor responsiveness. *Proc. Natl Acad. Sci. USA* **95**, 7157–7162 (1998).
9. Bohn, L. M. *et al.* μ -opioid receptor desensitization by β -arrestin-2 determines morphine tolerance but not dependence. *Nature* **408**, 720–723 (2000).
10. Potenza, M. N. & Nestler, E. J. Effects of RGS proteins on the functional response of the μ opioid receptor in a melanophore-based assay. *J. Pharmacol. Exp. Ther.* **291**, 482–491 (1999).
11. Gaudet, R., Savage, J. R., McLaughlin, J. N., Willardson, B. M. & Sigler, P. B. A molecular mechanism for the phosphorylation-dependent regulation of heterotrimeric G proteins by phosphoducin. *Mol. Cell* **3**, 649–660 (1999).
12. Berman, D. M. & Gilman, A. G. Mammalian RGS proteins: barbarians at the gate. *J. Biol. Chem.* **273**, 1269–1272 (1998).
13. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
14. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
15. Pearson, W. R. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* **11**, 635–650 (1991).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

Preparation of this report was supported by the National Institute on Drug Abuse.

Correspondence and requests for materials should be addressed to E.J.N. (e-mail: eric.nestler@utsouthwestern.edu).

A genomic view of immunology

Aude M. Fahrner*, J. Fernando Bazan†, Peter Papathanasiou*, Keats A. Nelms* & Christopher C. Goodnow*

*ACRF Genetics Laboratory and Medical Genome Centre, John Curtin School of Medical Research, Australian National University, Canberra 2601, Australia

†Department of Molecular Biology, DNAX Research Institute, 901 California Avenue, Palo Alto, California 94304-1104, USA

The outstanding problems facing immunology are whole system issues: curing allergic and autoimmune disease and developing vaccines to stimulate stronger immune responses against pathogenic organisms and cancer. We hope that the human genome sequence will reveal the molecular checks and balances that ensure both an effective immunogenic response against pathogenic microorganisms and a suitably tolerogenic response to self antigens and innocuous environmental antigens. Three synergistic approaches—sequence homology searches, messenger RNA expression profiling on microarrays, and mutagenesis in mice—provide the best opportunities to reveal, in the genome sequence, key proteins and pathways for targeting by new immunomodulatory treatments.

The human genome sequence contains a list of all the parts controlling beneficial and harmful immune responses. To identify these parts, the immediate hurdles of gene identification and genome assembly will need to be overcome. The ultimate challenge will be to bridge the gulf between a list of sequences and whole organism biology.

Predicting immune genes from sequence homology

The most immediate need is a list of all genes, exons and control elements in the human genome, but our experience is that the public domain data are still some way from this goal. To investigate the usefulness of searching the genome sequence on the basis of predicted protein homology alone (Fig. 1), we chose three examples of immunologically important proteins and asked how many more protein relatives could be found. Only the first of these examples could be searched with publicly available web-based databases. For details and advice about search options and databases, see Supplementary Information.

First, we looked for proteins of the tumour necrosis factor receptor (TNFR) family, members of which control proliferation or apoptosis of lymphocytes. Sequence motifs for the TNFR family have been assembled, for example in Interpro (see Supplementary Information). The search identified 21 of the 22 known TNFR family members¹, the exception being TNFR superfamily member 18. Five proteins did not immediately match known TNFR family proteins. One was the recently published TNFR family member TA α , and two ultimately corresponded to family members OPG and DR5. The two remaining proteins represented potentially novel family members: IGI_M1_ctg13384_53 was similar to OX40, and IGI_M1_ctg13980_78 was similar, but only over its amino-terminal half, to CD30. The fact that we found so few new members of the TNFR family may reflect either intense mining of expressed sequence tag (EST) databases or the incomplete assembly of raw genome data at present.

Second, we looked for members of the B7 family of costimulatory proteins. CD80 and CD86 (B7.1 and B7.2) are expressed on antigen-presenting cells. They engage CD28 and CD152 (CTLA-4) receptors on T cells to transmit either costimulatory or inhibitory signals to the T cell. Two homologues of CD80 and CD86 have recently been identified²: ICOSL, which binds to the T-cell costimulatory receptor ICOS (inducible costimulator); and B7H-1, which binds to the lymphocyte inhibitory receptor PD-1 (ref. 3). The B7 costimulatory proteins are not identified by specific motifs in the major protein databases, apart from being members of the immunoglobulin superfamily. Using PSI-BLAST searches (see Supplementary Information), we identified 21 proteins with homology to the B7 proteins.

Nine of these were members of the butyrophilin family.

Butyrophilin is a cell-surface protein, also found in breast milk, that has been identified as a B7 homologue⁴. Three of the proteins corresponded to signal regulatory protein (SIRP)- α -1 and one to SIRP- β -1. SIRP- α -1 is an inhibitory receptor, expressed by splenic macrophages, which binds to the CD47 self marker on erythrocytes and prevents their elimination; SIRP- β -1 seems to be involved in the activation of myeloid and dendritic cells.

We identified four other proteins corresponding to previously cloned genes: HHLA2, a human endogenous retrovirus sequence encoding a potentially secreted protein expressed in several tissues, including lymphocytes; MCAM, a melanoma adhesion protein apparently involved in tumour progression; CXADR, a surface molecule of unknown function; and VEJAM, a vascular endothelial junction-associated molecule, which may be involved in lymphocyte homing.

We also found four novel proteins (see Supplementary Information). Extrapolating from the known members, the less characterized proteins could be involved in the costimulation or modulation of lymphocytes, macrophages or other cell types.

The third homology search was performed on the basis of three-dimensional structure rather than amino-acid homology. Cytokines are an important class of immune regulators with protein folds that accommodate considerable sequence divergence⁵. Family relationships have often emerged only after the resolution of prototype cytokine folds; for example, cementing the distant similarity between fibroblast growth factors and interleukin-1 (IL-1)-like molecules⁶, or suggesting a link between TNF proteins and an extended family of complement C1q-like cytokines⁷. These findings often broaden our view of the evolution of cytokines as well as their biological functions. The challenge is to detect novel molecules with weak or unapparent ties to existing cytokine groups. We can do this best with computational techniques that are being used sensitively to annotate genome-derived sequences, and to fuse knowledge of the structural templates with sensitive sequence searching and prediction routines⁸.

We looked at the superfamily of haemopoietic cytokines, which is distinguished by a four-helix bundle fold that engages a special class of transmembrane receptor⁵. Although difficult to align by sequence similarity, the helical scaffolds of these cytokines reveal faint, subfamily-distinctive motifs when superimposed. Taking the best conserved 'D' helix alignment of a diverse series of IL-6-like cytokine structures (IL-6, granulocyte colony stimulating factor, ciliary neurotrophic factor (CNTF), oncostatin-M, leukaemia inhibitory factor and a carefully arrayed set of cardiotrophin-1, IL-11 and IL-12 sequences), we constructed weighted profiles, position-specific scoring matrices and hidden Markov models and used them iteratively to search both EST and genomic databases. These profiles collected all extant IL-6-type sequences, as well as a set of novel,

predicted open reading frames (ORFs) that were then used to clone their complete gene sequences. Among these orphan cytokines is a molecule that distantly resembles CNTF, called novel neurotrophin-1 or cardiotrophin-like cytokine—not surprisingly, this molecule signals by co-opting the CNTF receptor complex⁹.

Another outlier sequence resembles the p35 subunit of IL-12, and competes for binding to the p40 chain (creating a cytokine labelled as IL-23); it then binds to a receptor complex that includes elements of the IL-12 signalling machinery¹⁰. Thus, in cases where no sequence similarity is detectable—but secondary structure prediction indicates, for example, a compatible register of helices and loops—fold recognition or threading techniques can tease out a reliable alignment of a novel sequence with a helical cytokine template.

It is difficult to gauge how completely these searches scan the genome, because the genome sequence is not complete, and considerable work is still needed to identify correctly all protein-coding segments and to remove intronic sequences. It must also be remembered that to use protein homology queries at least one member of a family must already be known and must have a searchable motif.

Predicting immunological proteins from expression pattern

Sequence homology alone is a very incomplete predictor of genes in the immunological parts list, as many will not have known homologues and those that are revealed by this approach may not be expressed in the immune system. Expression of mRNAs or proteins in lymphoid cells is a stronger, complementary predictor (Fig. 1), particularly when the expression pattern changes during lymphocyte development or activation. Already, compilations of known genes and ESTs have been used to link sets of genes with differences in signalling brought about by tolerance or immunosuppression¹¹, and to link clusters of genes with different normal and neoplastic lymphoid cell types¹². Alignment of the human and mouse genome will illuminate regions of conservation, allowing oligonucleotide-based microarrays to assay all putative exons for transcription into mRNA in lymphoid cell types. This strategy holds the most

immediate promise for compiling a comprehensive parts list for the immune system, revealing differentially spliced forms as well as differences in overall mRNA abundance among immune cell subsets.

The genome sequence will also facilitate rapid identification of differentially expressed or modified proteins in immune cells or fluids by mass spectrometry and limited peptide microsequencing. This technique is limited by the need for better ways to resolve and quantify tens of thousands of proteins, which vary in abundance over many orders of magnitude.

Mutations and polymorphisms link genes and function

Showing that a variant gene alters an immunological trait provides the most certain way to bridge the gap between genome sequence and immune responses at the organismal level (Fig. 1). As many new genes are implicated in the immune system by expression profile and sequence homology, the chief bottleneck lies in testing the immunological effects of mutant alleles of these genes. Important functional data can be obtained relatively rapidly by *in vitro* assays and *in vivo* blocking with antibodies or Fc fusion proteins. There are nevertheless many examples where the function of a gene for the immune system as a whole was inaccurately extrapolated from such assays, compared to the role that was revealed in mice with loss-of-function mutations.

It is reasonable to propose that a systematic effort be mounted to knock out every immunologically expressed gene. This could be done by homologous recombination, but at an estimated cost of US\$800 million, assuming that there are 20,000 immune genes. Lexicon Inc. have developed a large panel of embryonic stem cell lines carrying individual genes inactivated and tagged by retroviral insertional mutagenesis which can be searched via the web (www.lexgen.com). The limitation of this method is that insertion is nonrandom. Of 67 B-cell activation genes identified by microarray profiling¹¹, one-third had retroviral insertions. For those that were targeted, there were usually multiple separate lines, often with insertions in the same location.

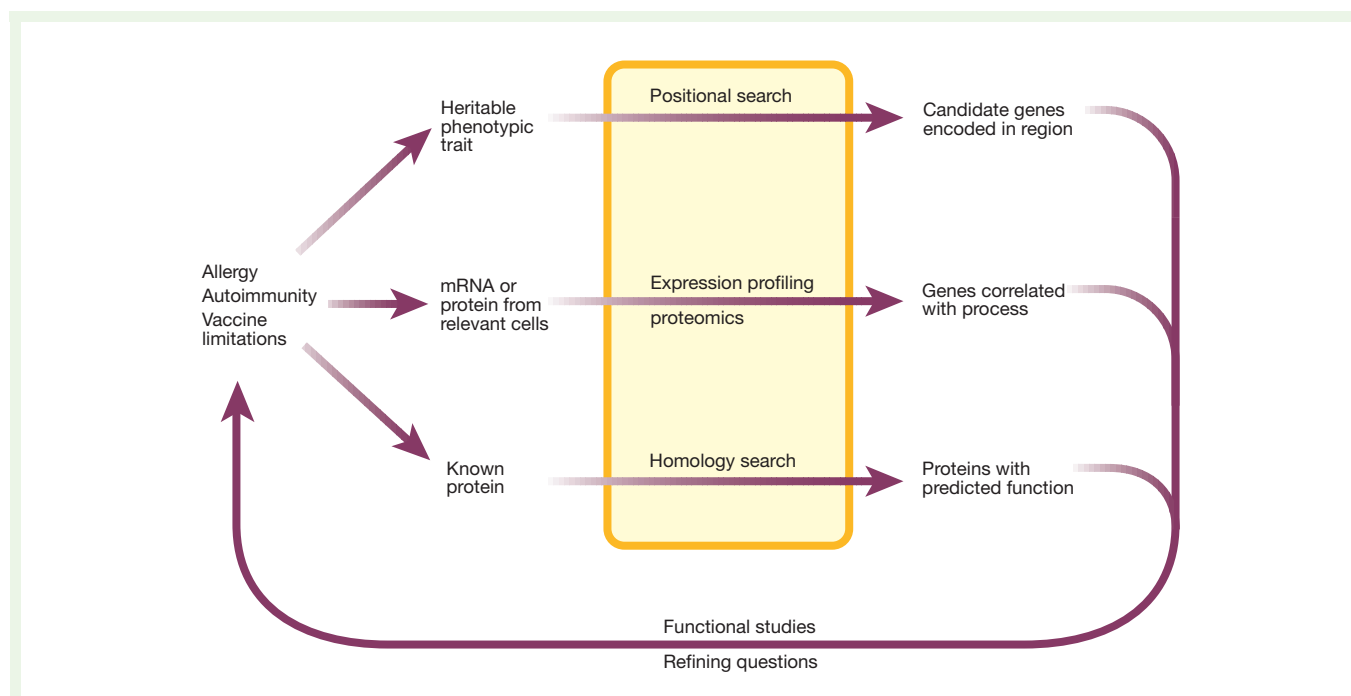


Figure 1 Impact of the human genome sequence on immunology. Three paths are shown for linking immune phenomena with individual genes; in each case, the human genome sequence has a central role.

In contrast to the 'reverse genetics' strategies above, which create mutants in predicted immunological genes using a bottom-up strategy, perhaps the greatest contribution made to immunology by the genome sequence lies in facilitating 'forward genetics' approaches. These start with an altered immunological trait and proceed from the top down to identify the causative gene. The trait may be a difference between strains of mice or a difference in human families. Research efforts can be focused on genes with key roles in immunological processes, regardless of sequence or expression pattern. A clear example is the identification of the mutated gene, AIRE, responsible for a mendelian syndrome of endocrine autoimmunity and candida infections (APECED)^{13,14}. The genome sequence allowed identification of candidate genes within the chromosomal interval to which the APECED mutation had been mapped. Resequencing of these candidates from affected individuals revealed the mutant AIRE gene. Our experience is that public genome databases have not quite reached the goal of laying out all candidate genes between two markers. Using a test search for an interval on chromosome 21, not all known markers could be located on the chromosome map, and there were gaps in gene annotation (see Supplementary Information).

With the mouse genome sequence to be completed in 2001, phenotype-driven gene identification strategies can now be performed on a large scale by genome-wide mutagenesis in the mouse. Use of the supermutagen ethylnitrosourea (ENU) to induce point mutations in a large fraction of genes, coupled with sensitive screens for immunological traits, is allowing identification of many key regulators of the immune system^{15,16}.

The complete list of parts provided by the human genome sequence will, on its own, not solve the major questions facing immunologists. Rather, it will form the basis of experiments designed to understand how the parts fit together to control immune responses. □

1. Screaton, G. & Xu, X. N. T cell life and death signalling via TNF-receptor family members. *Curr. Opin. Immunol.* **12**, 316–322 (2000).
2. Mueller, D. L. T cells: A proliferation of costimulatory molecules. *Curr. Biol.* **10**, R227–230 (2000).
3. Freeman, G. J. *et al.* Engagement of the PD-1 immunoinhibitory receptor by a novel B7 family member leads to negative regulation of lymphocyte activation. *J. Exp. Med.* **192**, 1027–1034 (2000).
4. Henry, J., Miller, M. M. & Pontarotti, P. Structure and evolution of the extended B7 family. *Immunol. Today* **20**, 285–288 (1999).
5. Sprang, S. R. & Barr, P. J. Cytokine structural taxonomy and mechanisms of receptor engagement. *Curr. Opin. Struct. Biol.* **3**, 815–827 (1993).
6. Zhang, J. D., Cousens, L. S., Barr, P. J. & Sprang, S. R. Three-dimensional structure of human basic fibroblast growth factor, a structural homolog of interleukin 1 beta. [published erratum appears in *Proc. Natl Acad. Sci. USA* **88**, 5477 (1991)]. *Proc. Natl Acad. Sci. USA* **88**, 3446–3450 (1991).
7. Shapiro, L. & Scherer, P. E. The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor. *Curr. Biol.* **8**, 335–338 (1998).
8. Fischer, D. & Eisenberg, D. Predicting structures for genome proteins. *Curr. Opin. Struct. Biol.* **9**, 208–211 (1999).
9. Elson, G. C. *et al.* CLF associates with CLC to form a functional heteromeric ligand for the CNTF receptor complex. *Nature Neurosci.* **3**, 867–872 (2000).
10. Oppmann, B., Bazan, J. F. & Kastelein, R. A. IL-23. *Immunity* (in the press).
11. Glynne, R. *et al.* How self-tolerance and the immunosuppressive drug FK506 prevent B-cell mitogenesis. *Nature* **403**, 672–676 (2000).
12. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
13. The Finnish-German APECED Consortium. An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. *Nature Genet.* **17**, 399–403 (1997).
14. Nagamine, K. *et al.* Positional cloning of the APECED gene. *Nature Genet.* **17**, 393–398 (1997).
15. Goodnow, C. C. *et al.* Mechanisms of self-tolerance and autoimmunity: From whole-animal phenotypes to molecular pathways. *Cold Spring Harb. Symp. Quant. Biol.* **64**, 313–322 (1999).
16. Justice, M. J. Capitalizing on large-scale mouse mutagenesis screens. *Nature Rev. Genet.* **1**, 109–115 (2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

The authors would like to thank Jason Bock and Greg Quinn for invaluable help with computer searches, and Iain Wilson for helpful discussions.

Correspondence and requests for materials should be addressed to A.M.F. (e-mail: aude.fahrer@anu.edu.au).

A genomic perspective on membrane compartment organization

Jason B. Bock, Hugo T. Matern, Andrew A. Peden & Richard H. Scheller

Howard Hughes Medical Institute, Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, California 94305-5426, USA

Now that whole genome sequences are available for many eukaryotic organisms from yeast to man, we can form broad hypotheses on the basis of the relative expansion of protein families. To investigate the molecular mechanisms responsible for the organization of membrane compartments, we identified members of the SNARE, coat complex, Rab and Sec1 protein families in four eukaryotic genomes. Of these families only the Rab family expanded from the unicellular yeast to the multicellular fly and worm. All families were expanded in humans, where we find 35 SNAREs, 60 Rabs and 53 coat complex subunits. In addition, we were able to resolve the SNARE class of proteins into four distinct subfamilies.

We compared families of proteins involved in vesicle trafficking from four organisms whose genome sequences are mostly complete: *Saccharomyces cerevisiae* (yeast)², *Drosophila melanogaster* (fly)^{3,4}, *Caenorhabditis elegans* (worm)^{5,6} and *Homo sapiens* (human) (Table 1). Although the genomes are at various stages of completion, we feel that these results are an early view of the genome, not a premature one.

Vesicle-trafficking pathways

A hallmark of eukaryotic cells is the ability to segregate biochemical reactions within membrane-bound organelles, such as the endoplasmic reticulum, Golgi complex, endosomes and lysosomes. These compartments contain both resident proteins, which define each compartment according to its biochemical function, and proteins transiently passing through on the way to their final destinations. Cargo is loaded into a newly forming vesicle, which buds from a donor organelle and is transported to an acceptor organelle. Here it docks and fuses, delivering its contents (Fig. 1). These distinct steps are present in all vesicle-trafficking pathways, whether moving a protein from the endoplasmic reticulum to the Golgi for glycosylation or from an endosome to a lysosome for degradation. A central issue in cell biology has been to understand how a cell carries out the various steps in the life cycle of a transport vesicle, and how specificity is conferred to the system, so that, for example, an endoplasmic reticulum-derived vesicle delivers its cargo to the Golgi and not the lysosome¹.

The protein families that mediate vesicle trafficking are conserved through phylogeny from yeast to man, as well as throughout the cell from the endoplasmic reticulum to the plasma membrane (see

Supplementary Information). Each vesicle-trafficking step involves a similar set of proteins drawn from these families. By having different members of these families localized to distinct membrane compartments, specificity is encoded in the machinery itself.

Genome analysis

The biogenesis of transport vesicles is initiated through the recruitment of large multi-subunit protein complexes termed coats⁷. These coats are central to both cargo selection and deformation of the lipid bilayer into a budding vesicle (Fig. 1, 1). Each coat complex is recruited to a distinct membrane compartment within the cell, indicating that individual coats are involved in specific transport steps. A coat complex is formed by the interaction of several different subunits. For example, an AP-1 adapter complex, involved with transport from the trans-Golgi network to endosomes, consists of four subunits (γ , $\beta 1$, $\mu 1$ and $\sigma 1$). Individual subunits carry out discrete functions within the complex, effecting the localization, regulation and affinity of the complex for cargo molecules.

Our analysis reveals that the number of coat complexes has increased only slightly from yeast, fly, and worm to human (from six to seven); however, the number of individual coat subunits has expanded significantly in humans (Table 1). This indicates that, instead of evolving new coat complexes for specialized trafficking steps and to accommodate increased cargo complexity, mammals have used a modular system where specificity can be achieved through subunit exchange. For example, the adaptor medium chains $\mu 1A$ and $\mu 1B$ are very similar, but only AP-1 adaptor complexes containing $\mu 1B$ can efficiently transport low-density lipoprotein (LDL) and transferrin receptors to the basolateral surface in polarized cells.

The next steps in the life cycle of a transport vesicle involve proteins of the Rab (Ypt in yeast) family. Rabs are a subfamily of the Ras superfamily of low-molecular-mass GTPases and cycle between a GTP, membrane-bound, active state and a GDP, cytosolic, inactive state. When bound to GTP, Rabs interact with a host of different proteins, loosely termed Rab effectors⁸. These Rab effectors perform diverse functions from vesicle budding to vesicle transport by way of the cytoskeleton, and vesicle docking at the target membrane (Fig. 1, 1–3). Different Rabs are localized on distinct vesicles and organelles, and are positioned to demarcate particular membranes. Thus, through their localization and GTP state, Rabs can recruit and/or activate their various effectors at the correct time and to the correct place, providing an element of regulation to the vesicle-trafficking machinery.

The number of Rabs in the four eukaryotic genomes analysed scaled roughly with the total number of predicted genes. The human

Table 1 Number of members of vesicle-trafficking families

	<i>S. cerevisiae</i>	<i>C. elegans</i>	<i>D. melanogaster</i>	<i>H. sapiens</i>
Coat complexes (subunits)	6 (31)	6 (29)	6 (29)	7 (53)
Rabs	11	29	26	60
SNAREs	21	23	20	35
Qa-SNAREs/syntaxins	7	9	7	12
Qb-SNAREs/SNAP Ns	5	7	5	9
Qc-SNAREs/SNAP Cs	6	4	5	8
R-SNAREs/VAMPs	5	6	5	9
Sec1s	4	6	5	7
Total predicted genes ^d	6,241	18,242	13,601	30,000–50,000

For a more extensive analysis of these protein families, including accession numbers and searching methodology, see Supplementary Information. SNARE helical definitions were determined by protein profiling. Some proteins (SNAP-25) contain two SNARE-coil domains and thus are counted twice in the coil subdivisions.

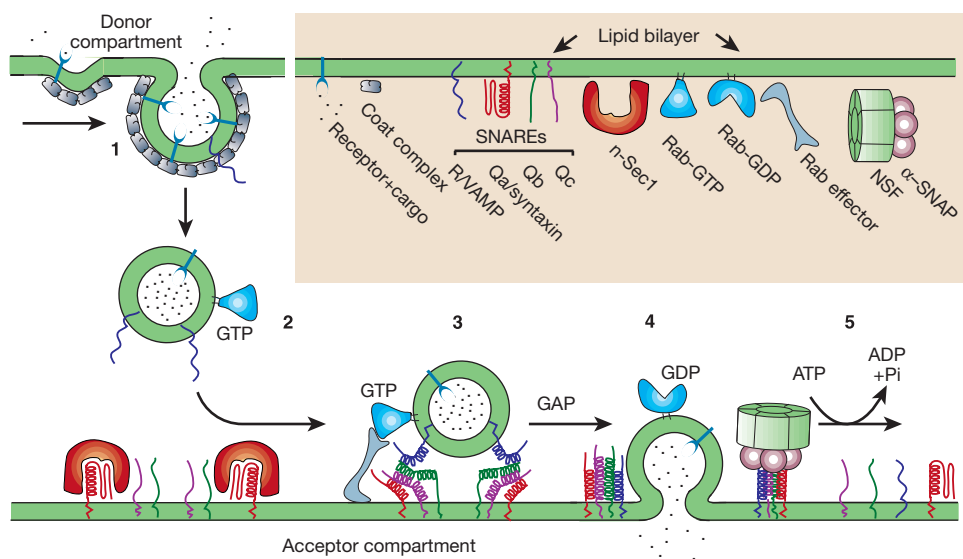


Figure 1 The life cycle of a transport vesicle. **1: Budding.** A small Arf/Sar GTPase is recruited to the donor membrane by its cognate guanine nucleotide exchange factor (see Supplementary Information). Next, incoming coat components are bound and cargo proteins diffuse into the budding site where they are trapped by their interaction with coat proteins or cargo receptors. The membrane forms a coated bud, giving rise to a vesicle. A GTPase activating protein (GAP) that is part of the coat stimulates GTPase activity, which leads to uncoating and release of the GTPase and coat proteins into the cytosol, where they can recycle. **2: Transport.** Vesicles are transported from their budding site to the acceptor compartment through association with cytoskeletal elements and transport motors (not shown). **3: Docking.** Vesicles are targeted to sites of fusion in an event that involves a GTP-bound Rab protein, elongated coiled-coil tethering proteins and other tethering factors (not shown). After the docking step, an R-SNARE (VAMP) assembles with a Qa-SNARE (syntaxin), a Qb-SNARE (SNAP N) and a Qc-SNARE (SNAP C) to form a parallel, four-helical bundle. Interactions of Sec1s with Qa-SNAREs are probably critical in forming the core fusion complex. **4: Fusion.** After nucleation of the ternary core fusion complex, further zippering of the parallel helices brings the vesicle and acceptor membrane close together, perhaps driving the fusion reaction. **5: Disassembly.** After fusion, Rab-GDP is released from the membrane by GDP-dissociation inhibitor (GDI) and recruited back to the donor membrane. The SNARE complex is disassembled by the ATPase *N*-ethylmaleimide-sensitive factor (NSF) and α -SNAP so that SNAREs can be recycled for further rounds of transport.

genome contains 60 Rabs, including 31 that were identified in this study (23 have known mammalian orthologues). As Rabs perform most of their functions by interacting with their effectors, until we have a more complete catalogue of those effectors and GDP/GTP exchange factors, we will not have a firm grasp on all of their functions. It has been difficult to identify Rab effectors because they share little, if any, sequence homology. This is a case where traditional genomics is of little help, but where the coming era of structural genomics may have an impact⁹. There is evidence that Rabs have both an essential role and a more subtle, regulatory role in vesicle trafficking. In yeast, Ypt1 (Rab1) is essential for trafficking from the endoplasmic reticulum to the Golgi apparatus. In contrast, Rab3A, present on synaptic vesicles, appears to have a regulatory role. Rab3A knockout mice are viable, but certain synapses in their brains fail to undergo a type of plasticity called long-term potentiation. Thus, whereas some Rabs appear to be essential parts of core vesicle-trafficking machinery, others appear to act more as regulators of the process. Perhaps an expansion in the regulatory role of Rabs was necessary to coordinate cell physiology with membrane trafficking in complex, multicellular organisms.

The final step in vesicle trafficking is the fusion of a vesicle with its target membrane, believed to be mediated by a family of proteins termed SNAREs¹. SNAREs are integral membrane proteins present on both vesicle and target membranes, and can form very stable complexes. The formation of a SNARE complex pulls the vesicle and target membrane together and may provide the energy to drive fusion of the lipid bilayers. It is proposed that a unique SNARE complex, consisting of four helices, is formed at each fusion site. For example, the fusion of synaptic vesicles with the presynaptic nerve terminal is mediated by the formation of a complex comprising one helix each from syntaxin 1A and VAMP-2 and two helices from

SNAP-25. These four helices are prototypes that can be used to define subclasses of SNAREs on the basis of protein profiling (see Table 1 and Supplementary Information). Different members of the SNARE families are localized to distinct membrane compartments throughout eukaryotic cells and thus are positioned to form specific complexes and to enhance the fidelity of vesicle trafficking at the final step.

The SNARE family has remained mostly unchanged in yeast, flies and worms, but has increased significantly in humans to 35 members. This leads to the conclusion that multicellular organisms do not have an inherently more complex secretory pathway, and that a set of core SNAREs is sufficient to mediate most intracellular vesicle fusion events. Mammals appear to have taken multicellular specialization to another level, with differential expression of more SNARE proteins. For example, syntaxin 17 is abundantly expressed in steroidogenic tissues, and is localized to the smooth endoplasmic reticulum. Neurotransmission is a specialized form of exocytosis mediated by syntaxins 1A and B, which are greatly enriched in neurons. So, although multicellular organisms clearly can function with a core set of SNAREs, use of additional SNAREs may result in further tissue-specific specialization of membrane trafficking.

A protein family that interacts directly with the syntaxin subfamily of SNAREs is classified by the prototypical member Sec1 (also known as mUNC-18). Sec1s are cytosolic proteins, which are peripherally associated with membranes in part through their interactions with syntaxins. Sec1 is essential for vesicle trafficking; a knockout of *SEC1* in yeast abolishes exocytosis. The ratio of Sec1s to syntaxins in the genomes analysed remains roughly constant, supporting the idea that Sec1s act primarily through syntaxins. They are probably chaperones, putting syntaxins into conformations that are required for interactions with other SNAREs. As with

the SNAREs, the increase in Sec1 genes can be attributed to specialization of trafficking events. For example, the Sec1 family member mUNC-18c interacts with syntaxin 4 and is involved in the translocation of glucose transporters to the cell surface in response to insulin.

Conclusions

Analysis of the extent of four families of proteins involved in vesicle trafficking leads to several conclusions. The evolutionary leap from single to multicellular organisms did not require an increase in the core machinery (coats, SNAREs, Sec1s) underlying the transport vesicle life cycle. The Rab family did expand significantly from yeast to worms and flies, implying that multicellular organisms exert more regulation over vesicle-trafficking pathways. The evolutionary jump to mammals saw a significant increase in all of the families examined. This implies that mammals orchestrate the complexities of multicellular physiology through not only more finely tuned regulation, but also tissue-specific specialization of the core trafficking machinery. This is perhaps most clearly illustrated in the brain, where memories are at least partially encoded through regulation of specialized membrane trafficking proteins most abundantly expressed in neurons. □

1. Lin, R. C. & Scheller, R. H. Mechanisms of synaptic vesicle exocytosis. *Annu. Rev. Cell Dev. Biol.* **16**, 19–49 (2000).
2. Goffeau, A. *et al.* The Yeast Genome Directory. *Nature* **387**, (suppl.), 1–105 (1997).
3. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
4. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
5. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology [published errata appear in *Science* **283**, 35 (1999), *Science* **283**, 2103 (1999) and *Science* **285**, 1493 (1999)]. *Science* **282**, 2012–2018 (1998).
6. Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J. & Spieth, J. Wormbase: network access to the genome and biology of *C. elegans*. *Nucleic Acids Res.* (in the press).
7. Hirst, J. & Robinson, M. S. Clathrin and adaptors. *Biochim. Biophys. Acta* **1404**, 173–193 (1998).
8. Novick, P. & Zerial, M. The diversity of Rab proteins in vesicle transport. *Curr. Opin. Cell Biol.* **9**, 496–504 (1997).
9. Skolnick, J., Fetrow, J. S. & Kolinski, A. Structural genomics and its importance for gene function analysis. *Nature Biotechnol.* **18**, 283–287 (2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) and as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank E. Birney and his colleagues at EBI for their help and patience with the human genome dataset, L. Kozar at Stanford and J. Chesnick at the Genetics Computer Group for technical support, and M. Murthy, S. Scales and F. Assaad for discussions.

Correspondence and requests for materials should be addressed to R.H.S. (e-mail: scheller@cmgm.stanford.edu).

Genomics, the cytoskeleton and motility

Thomas D. Pollard

Structural Biology Laboratory, Salk Institute for Biological Studies, 10010 N. Torrey Pines Road, La Jolla, California 92037, USA

The draft human genome sequence is an important step in cataloguing the molecular hardware that supports the processes of life. Here I look at what we have learned from the draft sequence about our cytoskeletal and motility systems. Most cytoskeletal and motility proteins were discovered previously by biochemical isolation, traditional cloning methods or random sequences of complementary DNAs. The ongoing challenges of assembling and annotating genes for motor proteins with long, fragmented coding sequences emphasize the importance of expert knowledge of related proteins and confirmatory evidence from cDNA sequences.

Humans and other higher animals have three cytoskeletal systems: actin filaments, microtubules and intermediate filaments¹. The myosin family of molecular motors pull on actin filaments or move cargo along them. Dynein and kinesin motors move microtubules or move cargo along microtubules.

How many genes?

Traditional biochemical and genetic methods have identified many protein components of these three systems: 6 mammalian actins and more than 70 families of actin-binding proteins; about 6 vertebrate α -tubulins and β -tubulins (forming the dimeric subunit of microtubules) and about a dozen families of microtubule-binding proteins; and about 31 human intermediate filament proteins and 5 families of associated proteins. Most families of proteins that bind one of the cytoskeletal polymers consist of several isoforms, resulting from multiple genes, alternative splicing or both. Divergent actins called actin-related proteins or Arps² have special functions and contribute to the diversity of the system.

The draft genome sequence might have unveiled a cornucopia of new cytoskeletal genes. Instead, it appears to confirm that most of these genes were found by traditional methods. For example, the known actin family included six functional actin genes, more than twenty pseudogenes and six families of Arps encoded by nine genes (Table 1). The draft genome locates many of the pseudogenes and reveals at least fourteen new genes: seven highly divergent actin genes and seven new Arp genes. Work is required to establish whether these genes are functional. More genes may emerge, as two known actin genes are not recognized in the draft sequence.

Genes for complex, multi-domain proteins such as the motor proteins myosin and kinesin are more challenging to assemble than genes for small, single-domain proteins such as profilin and actin (Fig. 1). However, researchers found virtually all of the genes for myosin and kinesin using traditional cloning methods and searches of cDNA databases. The current annotation of the draft human genome sequence includes most of those genes, but many are represented as fragments that have not yet been assembled into full-length coding sequences. The genes are huge, fragmented by introns, and include multiple non-homologous domains outside the core energy-transducing domains.

Myosin and kinesin researchers found about 40 myosin genes³ and 40 kinesin genes (R. Freedman and K. Wood, personal communication). These were found initially by biochemical isolation, directed cloning and characterization. Searches of emerging cDNA and genomic databases completed the inventories of full-length genes. Myosin and kinesin each have their own defining catalytic domain, the major features of which remain in protein sequences even after many rounds of gene duplication and considerable sequence divergence. Sequence comparisons readily identify these catalytic domains, but finding the rest of these large genes is

challenging. Most kinesins and myosins have a large carboxy-terminal tail consisting of many divergent domains (Fig. 1). Some have amino-terminal domains as well, and many introns fragment the genes. Nevertheless, an experienced worker with in-depth knowledge of the gene families can usually piece together complete genes by starting with a homologous catalytic domain and searching cDNA and genomic sequences in public databases for the flanking sequences.

Given the draft human sequence, annotators set out to predict coding sequences globally. The gene assembly procedures identified coding sequences for parts or all of many homologous catalytic domains of the motor proteins. A search of the contracted protein dataset of 17 July 2000 with the catalytic domain of a human cytoplasmic myosin returned about 80 coding sequences for myosins, but few correct, full-length coding sequences for myosin genes that had not already been defined by traditional methods. Few known myosin genes were missed, but owing to the fragmentary nature of the sequences on the current hit list, some hits are parts of the same gene. Thus the final number of myosin genes will be fewer than the current hits. The situation is similar for kinesins. Therefore, current estimates of the number of human genes for complex, multidomain proteins are likely to be inflated to some extent and must be considered to be provisional.

Table 1 Human actin and Arp (actin-related protein) gene families

Gene type	Number, genes (pseudogenes) and new genes (in bold)	Chromosome locations
α -actin skeletal muscle	1	1
α -actin cardiac muscle	1	Unknown
α -actin smooth muscle	1	10
β -actin cytoplasmic	1 (22–23)	7 (1, 1, 2, 2, 3, 3, 5, 5, 5, 5, 5, 6, 16, 16, 16, 17, 5 unknown)
γ -actin cytoplasmic	1 (6)	17 (1, 3, 11, 19, X, unknown)
γ -actin smooth muscle	1	2
Divergent actins	7	3, 11, 18, 19, 20, 2 unknown
Arp1 α , β	2	10, unknown
Arp1-like	1	2
Arp2	1	2
Arp2-like	1	3
Arp3 α , β	2	2, 7
Arp3-like	4	4, 4, X, unknown
Arp6/BAF53b	2	2, unknown
Arp7A, B	2	2, unknown
Arp11	1 + 1	2, unknown
Total genes (pseudogenes)	30 (28–29)	

References for known genes: 7, 8, 9. This inventory was made by searching the International Protein Index (<http://www.ensembl.org/IPI>) of 17 July 2000 with the amino-acid sequences of human skeletal muscle α -actin and Arp2. The searches returned identical lists of genes but with different rankings. The proteins were classified using BLAST searches of the NCBI protein database with sequences of these hits to identify related genes. Chromosome locations were determined at the UCSC website (genome.ucsc.edu). Additional pseudogenes have been mapped by FISH: β -actin (chromosomes 15 and 18); and γ -actin (chromosomes 6 and 20)¹⁰. 'Divergent actins' are related more closely to actins in other species (fungi, plants, protozoa) than vertebrate actins.

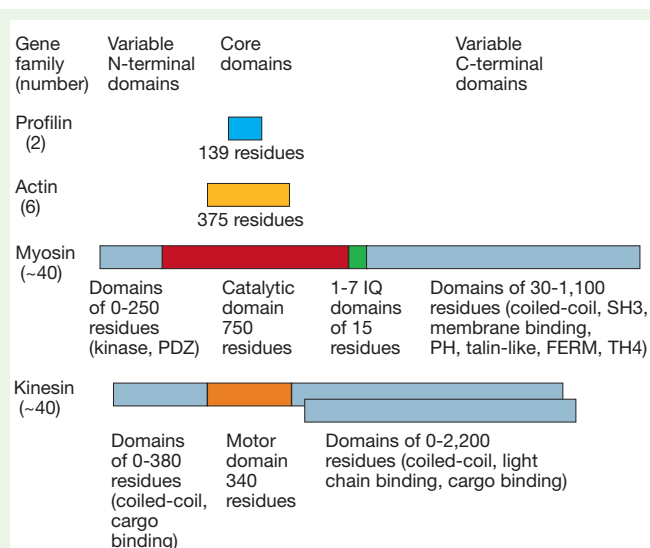


Figure 1 Domain maps of four types of cytoskeletal protein. Variable N- and C-terminal domains are light blue; other colours represent core domains. Neither profilin nor actin has accessory domains. Myosin and kinesin genes have a range of variable domains at the N- and/or C-termini of the core 'catalytic' and 'motor' domains. Myosins have 1–7 IQ domains, which bind light chains or calmodulin.

The WASp/Scar family represents the middle ground. These signal-transducing proteins regulate actin filament nucleation by the Arp2/3 complex⁴, and are less well characterized than myosins or kinesins. They are not large, but their multiple, incompletely conserved domains present challenges for gene assembly. All have WH2 domains (which bind an actin monomer), A domains (which bind Arp2/3) and proline-rich sequences (which bind profilin and SH3 domains). WASPs, but not Scars, also have WH1 domains (which bind WIP/verprolin) and GTPase-binding domains (which bind Cdc42). Traditional methods revealed human genes for WASp, N-WASP and three Scars (also called WAVE). The provisional human protein dataset includes several candidates for new WASp/Scar proteins, identified by sequence similarity to WH1, WH2/A or A domains (W.-L. Lee, personal communication). RNB6 appears to have a WH1 domain, but no WH2 or A domains. IGI_M1_ctg18730_17 and Q9Y6W5 have WH2 and A domains, but are gene fragments requiring further assembly, verification with cDNAs and testing for relevant activities.

These examples illustrate the importance of individual effort by biological specialists in completing the annotation of the human genome. Given inevitable sequencing errors and huge introns, accurate gene assembly from genomic sequences requires in-depth knowledge of related genes. Improving the completeness of collections of full-length cDNA sequences will be valuable in validating the annotation of genome sequences. The most constructive approach might be to develop guidelines and tools for annotating genes and gene families and then farm out most of the work to experts on each biological process, as suggested by Brinkman *et al.*⁵. This is one case where small science will yield a better product than the industrial approach required for sequencing.

Impact

Once completed, will the inventory of genes aid or distract from the search for general principles? For 50 years the model for understanding of biology has been to study a model system in depth and then to extrapolate its principles to related physiological systems.

However, I know of no protein that has been understood in any depth without a decade or more of intense effort by laboratories devoted to determining its structure, interactions with partner molecules and roles in cellular physiology. Given finite resources, scientists cannot analyse every human gene product in such detail. I fear that the temptation to analyse full genome sequences will prove irresistible, perhaps even delaying the laboratory work required to complete the mechanistic analysis needed to understand physiology.

The complete genetic inventory will advance our understanding of disease. Although estimates vary, most human genes may contribute to disease. For example, genetic variants in each major contractile protein cause dysfunction of the human heart⁶. The same must be true in other organs. Thirty years ago the conditions caused by these mutations were called 'idiopathic' cardiomyopathies, and as recently as 1986 the idea that amino-acid substitutions in contractile proteins cause cardiomyopathies was still speculative.

The inventory will also reveal useful targets for development of new therapies. We now know, in principle, most of the potential drug targets that are available to treat human diseases. Of course, years of validation of these targets lies ahead, but the genome sequence provides limits. In the case of the cytoskeleton, this opportunity is untapped. Other than cancer drugs that bind microtubules (vinca alkaloids and taxol), no drugs in clinical use bind proteins of the cytoskeleton or molecular motors that operate our cardiovascular and musculo-skeletal systems.

The genetic inventory is essential to appreciate the complexity of the cytoskeleton, associated molecular motors and other systems. However, I doubt that the inventory of genes will provide much insight into molecular mechanisms, as even the simplest protein is multifaceted and has a complex mechanism of action. Genomics may help to identify networks of interacting proteins, but understanding how protein networks function requires details about the dynamics of the reaction pathways as well as the concentration and cellular localization of each component. Only continued work on atomic structures and the biophysics of molecular interactions and reactions, along with genetic or pharmacological tests for physiological functions, will reveal the mechanisms required to understand the essence of physiology and pathology. □

- Kreis, T. & Vale, R. (eds) *Guidebook to the Cytoskeletal and Motor Proteins* 2nd edn (Oxford Univ. Press, Oxford, New York, 1999).
- Schroer, T. A. *et al.* Actin-related protein nomenclature and classification. *J. Cell Biol.* **127**, 1777–1778 (1994).
- Berg, J. S., Powell, B. C. & Cheney, R. E. A millennial census of the myosin superfamily. *Mol. Biol. Cell* (in the press).
- Higgs, H. N. & Pollard, T. D. Regulation of actin polymerization by Arp2/3 complex and WASp/Scar proteins. *J. Biol. Chem.* **274**, 32531–32534 (1999).
- Brinkman, F. S. L., Hancock, R. E. W. & Stover, C. K. Sequencing solution: use volunteer annotators organized via Internet. *Nature* **406**, 933 (2000).
- Towbin, J. A. The role of cytoskeletal proteins in cardiomyopathies. *Curr. Opin. Cell Biol.* **10**, 131–139 (1998).
- Gunning, P. *et al.* Isolation and characterization of full length cDNA clones for human α -, β - and γ -actin mRNAs: skeletal but not cytoplasmic actins have an amino-terminal cysteine that is subsequently removed. *Mol. Cell. Biol.* **3**, 787–795 (1983).
- Ng, S.-Y. *et al.* Evolution of the functional human β -actin gene and its multi-pseudogene family: conservation of noncoding regions and chromosomal dispersion of pseudogenes. *Mol. Cell. Biol.* **5**, 2720–2732 (1985).
- Schafer, D. A. & Schroer, T. A. Actin-related proteins. *Annu. Rev. Cell. Dev. Biol.* **15**, 341–363 (1999).
- Ueyama, H., Inazawa, J., Nishino, H., Ohkubo, I. & Miwa, T. FISH localization of human cytoplasmic actin genes ACTB to 7p22 and ACTG1 to 17q25 and characterization of related pseudogenes. *Cytogenet. Cell Genet.* **74**, 221–224 (1996).

Acknowledgements

I thank E. Birney and A. Kasprzyk (European Bioinformatics Institute) for searching the human protein datasets for cytoskeletal proteins, and J. Berg and R. E. Cheney (Univ. North Carolina) and W.-L. Lee (Washington Univ.) for analysing these datasets.

Correspondence should be addressed to T.D.P. (e-mail: pollard@salk.edu).

Can sequencing shed light on cell cycling?

Andrew W. Murray*† & Debora Marks‡

*Center for Genomics Research, and †Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA

‡BCMP and Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA

Every organism must have cells that can replicate indefinitely. Can the draft human genome sequence tell us how the cell cycle works and how it evolved? We studied two protein families—the cyclins and their partners the cyclin-dependent kinases (Cdks)—and a conserved regulatory circuit, the spindle checkpoint. Disappointingly, we discovered a few novel cyclins and no new Cdks or components of the spindle checkpoint, and could shed little light on the organization of the cell cycle.

Oscillations in the activity of the Cdk family of protein kinases drive the eukaryotic cell cycle. These enzymes consist of a catalytic (Cdk) and a regulatory subunit (cyclin)¹. Their activity is regulated by phosphorylation of Cdks, binding of stoichiometric kinase inhibitors and regulated proteolysis of the cyclins. Cdk activity is regulated by events inside and outside the cell and controls DNA replication, chromosome segregation and cell division. Despite wide variation in the speed and appearance of the cell cycle, the biochemical engine that regulates it has been remarkably conserved.

Cdks and cyclins

The Cdk/cyclin family contains five Cdks (Cdk1, 2, 3, 4 and 6) and four cyclin classes (cyclin A, B, D and E) that regulate the cell cycle; three Cdks (Cdk7, 8 and 9) and three cyclin classes (cyclin C, H and T) that regulate transcription; and other Cdk and cyclin classes with less defined roles (see Supplementary Information). We searched the predicted human proteins for Cdks and cyclins. We found no new Cdks. This is not surprising, as many labs have used polymerase chain reaction from conserved sequences to search for Cdks. Their success indicates that all the members of a highly conserved, well studied family are likely to be known before a genome is fully sequenced.

The draft human genome does answer one question about Cdk activation. Activating the cell-cycle Cdks requires phosphorylation of a conserved threonine as well as binding of cyclin. In mammals, the activating kinase is another Cdk/cyclin complex (Cdk7/cyclin H), in budding yeast it is a non-Cdk kinase (Cak1)¹, and in fission yeast both types of kinase are present and can activate Cdk/cyclin

complexes². There is no homologue of Cak1 in the human genome, suggesting that humans and budding yeast use different kinases for the same reaction. It appears that this step is not an important way of regulating Cdk activity, as no physiological regulation of Cdk7/cyclin H or Cak1 has been found.

The sequences of the cyclins are less well conserved than those of the Cdks. Cyclins have been found by three approaches: characterizing proteins that regulate processes such as the cell cycle or transcription; sequencing genes that are regulated in interesting ways; and as components of whole genome sequences. We searched the predicted proteins of the human genome for homology to known cyclins from humans and yeasts, focusing on genes that were annotated as novel. These candidates were compared against several databases, eliminating those that were fragments of known human cyclins. The remainder included the human homologue of chicken cyclin B3 and three novel cyclins. First, cyclin P, whose sequence is deduced from a combination of gene prediction on the genomic sequence and the sequence of a complementary DNA clone. It is related to but distinct from the A and B cyclins (see Fig. 1a and Supplementary Information). Second, a new cyclin, cyclin M, which appears to be most closely related to cyclin L (Fig. 1b). Its biological function is unknown, but it is related to cyclins that regulate transcription. Third, a protein previously reported as uracil-DNA glycosylase 2, whose similarity to existing cyclins has been noted³. Because a recombinant protein has not been shown to have glycosylase activity, this protein may be a novel glycosylase-associated cyclin rather than a catalytically active glycosylase. We suggest that this protein be called cyclin O (Fig. 1a).

We also identified possible vertebrate homologues of the large Pcl



Figure 1 Sequence alignments. **a**, Alignments of the most conserved regions of the cyclin box for A and B type cyclins. The alignments of cyclin O (Genbank accession no. NM_021147) and P (IPI protein set¹⁰ accession number IGI_M1_ctg17876_3) are shown beneath. Green letters show residues conserved in vertebrate A and B type cyclins. **b**, Alignment of the maximally conserved region of the transcriptional cyclins. Residues that are identical or conservative substitutions between cyclin L1 and cyclin M (IPI protein set¹⁰ accession number IGI_M1_ctg16669_8) are shown in red, as are residues that also show conservation in cyclin I and K. IPI, <http://www.ensemble.org/IPI/>

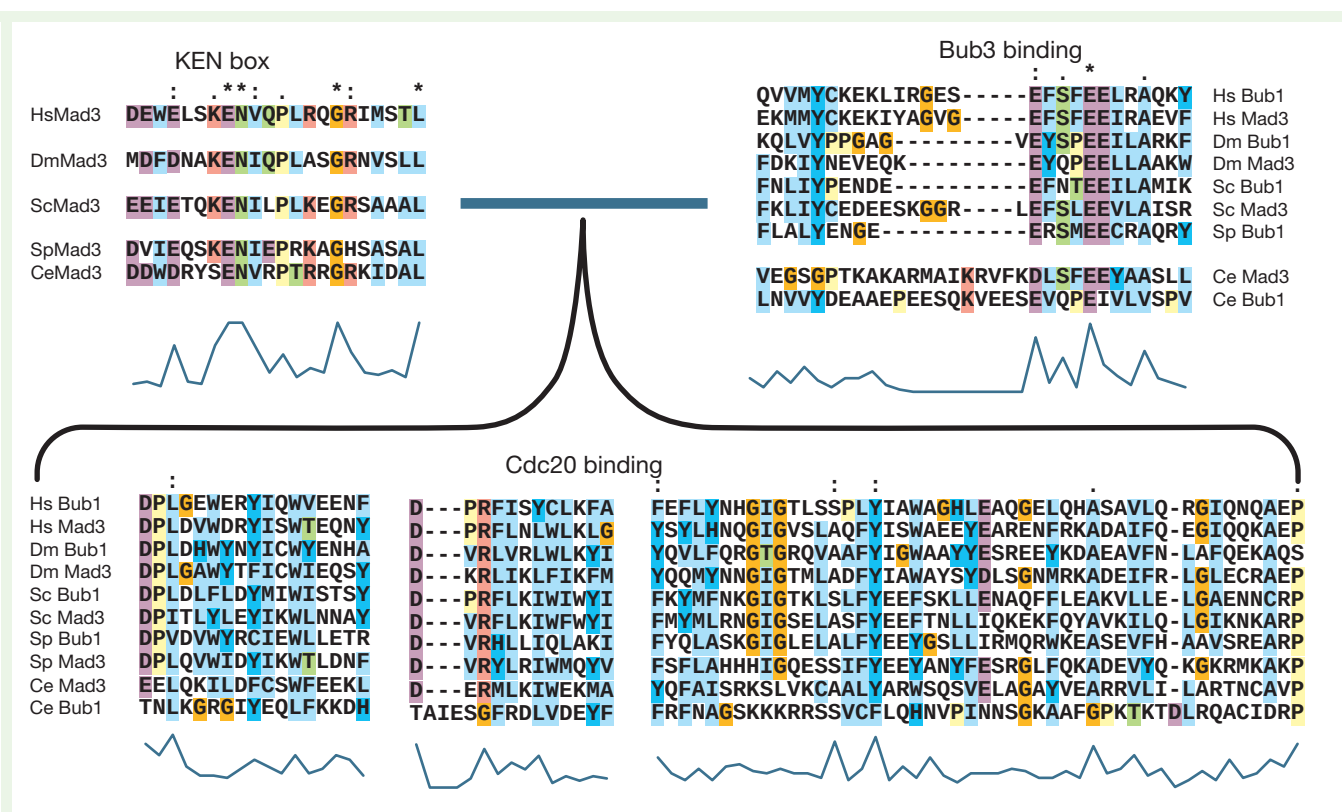


Figure 2 Sequence alignment, created using Clustal X, of portions of the Bub1 and Mad3 proteins. Includes the KEN box uniquely present in the N termini of Mad3 proteins, three stretches of conserved amino acids in the region that binds Cdc20, and conserved sequence in the region that binds to Bub3. Note the divergence of the worm sequences from those of the other eukaryotes. Note that the *Drosophila* Mad3 protein is labelled as Bub1 in databases and vice versa.

family of yeast cyclins⁴. A small set of related human proteins encodes proteins with very strong homology to fly and worm proteins, and the proteins from all three organisms show weak homology to the yeast Pcl proteins (see Supplementary Information), but it is unknown whether this homology has biological significance.

Sequence analysis can only hint at the functions and Cdk partners of these cyclins. A first hypothesis is that the novel cyclins will have similar functions to their closest relatives; the closer the relationship, the more likely this is to be true. The structure of one Cdk/cyclin complex is known⁵, so it should be possible to model the pairwise interactions of all Cdks with all cyclins and deduce which Cdks partner a given cyclin by looking for pairs whose hypothetical interface lacks either steric clashes or unfilled space.

The spindle checkpoint

The spindle checkpoint monitors the kinetochore, the proteinaceous complex that assembles on the centromeric DNA and attaches chromosomes to the microtubules of the spindle. Kinetochore that are not correctly attached to microtubules recruit the components of the checkpoint, initiating a signalling pathway that inhibits the anaphase-promoting complex (the enzyme that triggers chromosome segregation and the exit from mitosis⁶). The checkpoint proteins and pathway have been conserved during evolution, and mutations in two of the proteins have been implicated in the chromosomal instability that is widespread in human cancers⁷.

We examined the genomes of completely sequenced organisms for homologues of five known checkpoint genes (*Mad1–3*, *Bub1*, *Bub3*). All the organisms contain a single copy of each gene, but there were some surprises. The first is that the checkpoint proteins in worms have diverged more from those of yeasts, flies and mammals than the mammalian or fly proteins have diverged from

those of yeasts. In the worm, protein motifs that are conserved among other organisms have diverged or been lost altogether (Fig. 2). We speculate that this deviation reflects the unusual behaviour of worm chromosomes. In most eukaryotes, from yeast to human, each chromosome has only one functional kinetochore; worms have an elongated array of kinetochores that spans the length of the chromosome. Adapting to this different architecture may have imposed special challenges on the spindle checkpoint, causing its components to evolve more rapidly in the nematode lineage.

Correctly identifying two of the checkpoint proteins, Mad3 and Bub1, was remarkably difficult. In budding yeast, where the proteins were discovered, they share two regions of homology that allow them to interact with Bub3 and Cdc20 (the target to which the checkpoint binds when it arrests the cell cycle), but Bub1 has a carboxy-terminal protein kinase domain that Mad3 lacks. Other organisms contain a sequence with a similar organization to Bub1, as well as a related protein that can be as large as Bub1 and carry a protein kinase domain (mammals, fly and worm) or be even shorter than the budding yeast protein (fission yeast).

Which protein is Bub1 and which Mad3? Attempts to deduce the overall phylogeny of these proteins give confusing results, as the most conserved regions of the amino termini of the budding yeast Bub1 and Mad3 proteins are more closely related to each other than they are to either protein in fission yeast (Figs 2, 3). This may indicate that the N termini have co-evolved with their binding partners, Cdc20 and Bub3. Despite this overall similarity, there is one distinguishing feature. In every organism, only one of the paired proteins contains a conserved sequence near the N terminus that matches the KEN box that targets proteins for destruction at the end of anaphase⁸. This is an attractive finding, as once anaphase begins, the tension at the kinetochores starts to fall. Destroying one of the checkpoint components would eliminate the possibility that the

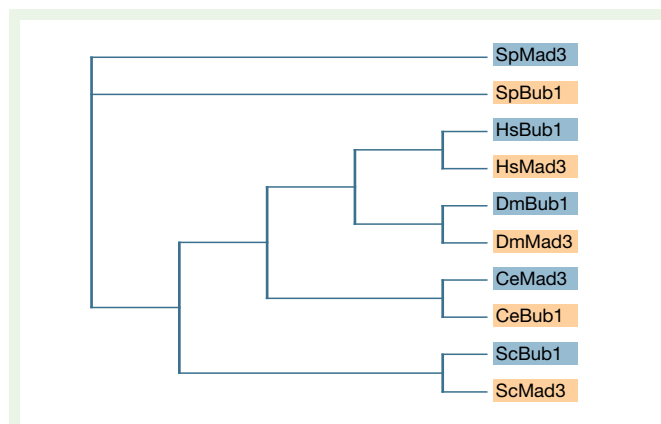


Figure 3 Phylogeny of the conserved regions of the N terminal portion of the Bub1 and Mad3 proteins. Sequences were aligned and trees created with Clustal X (<http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/>). Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Sc, *Saccharomyces cerevisiae*; Sp, *Schizosaccharomyces pombe*; Ce, *Caenorhabditis elegans*.

checkpoint could be inadvertently reactivated, impeding the exit from mitosis. In both yeasts, this motif is in the shorter, kinase-deficient protein, suggesting that the KEN box identifies the Mad3 homologues in all eukaryotes.

This comparison illustrates two points. First, phylogenetic analysis on different regions of the same class of proteins can produce different conclusions about the relationships among the protein family, reflecting the different evolutionary pressures faced by different parts of the protein. Second, there can be large-scale reorganizations of components of a conserved pathway. In the case of the spindle checkpoint and Mad3, the functional consequences of these changes are unclear.

Conclusions

Genome sequencing has revealed surprisingly little about the cell cycle. The two main conclusions of comparative analysis were

drawn well before the first eukaryotic genome was sequenced: the machinery that regulates the cell cycle has been highly conserved in eukaryotic evolution, and the size of protein families such as the Cdks and cyclins has expanded as organisms have become bigger and acquired more cell types. Comparing sequenced genomes has strengthened but not altered these conclusions. Disappointingly, comparing fully sequenced genomes does not explain how differences have evolved between the cell cycles of different organisms. For example, a conserved pathway (the mitotic exit network) is required for cytokinesis in budding and fission yeast, but is required to complete mitosis in budding but not in fission yeast⁹. Men live, and must avoid cancer, for thirty times as long as mice, and it is much easier to induce mouse cells to proliferate indefinitely *in vitro*. Can comparing the sequence and expression of two genomes reveal the source of these differences? Getting a positive answer to such questions will require a marked improvement in our ability to turn raw sequence data into biological knowledge. □

1. Morgan, D. O. Cyclin-dependent kinases: engines, clocks, and microprocessors. *Annu. Rev. Cell. Dev. Biol.* **13**, 261–291 (1997).
2. Lee, K. M., Saiz, J. E., Barton, W. A. & Fisher, R. P. Cdc2 activation in fission yeast depends on Msc6 and Csk1, two partially redundant Cdk-activating kinases (CAKs). *Curr. Biol.* **9**, 441–444 (1999).
3. Muller, S. J. & Caradonna, S. Cell cycle regulation of a human cyclin-like gene encoding uracil-DNA glycosylase. *J. Biol. Chem.* **268**, 1310–1319 (1993).
4. Moffat, J., Huang, D. & Andrews, B. Functions of Pho85 cyclin-dependent kinases in budding yeast. *Prog. Cell Cycle Res.* **4**, 97–106 (2000).
5. Jeffrey, P. D. *et al.* Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature* **376**, 313–320 (1995).
6. Amon, A. The spindle checkpoint. *Curr. Opin. Genet. Dev.* **9**, 69–75 (1999).
7. Cahill, D. P. *et al.* Mutations of mitotic checkpoint genes in human cancers. *Nature* **392**, 300–303 (1998).
8. Pfeiffer, C. M. & Kirschner, M. W. The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. *Genes Dev.* **14**, 655–665 (2000).
9. Jaspersen, S. L., Charles, J. F., Tinker-Kulberg, R. L. & Morgan, D. O. A late mitotic regulatory network controlling cyclin destruction in *Saccharomyces cerevisiae*. *Mol. Biol. Cell* **9**, 2803–2817 (1998).
10. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Correspondence and requests for materials should be addressed to A. M. (e-mail: amurray@mcb.harvard.edu).

Evolutionary analyses of the human genome

Wen-Hsiung Li, Zhenglong Gu, Haidong Wang & Anton Nekrutenko

Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637, USA

The completion of the human genome will greatly accelerate the development of a new branch of science—evolutionary genomics. We can now directly address important questions about the evolutionary history of human genes and their regulatory sequences. Computational analyses of the human genome will reveal the number of genes and repetitive elements, the extent of gene duplication and compositional heterogeneity in the human genome, and the extent of domain shuffling and domain sharing among proteins. Here we present some first glimpses of these features.

We have analysed the draft human genome sequence for data related to evolutionary genomics. Our investigations reveal new information about repetitive elements, domain sharing and conservation, and gene duplication in the human genome (for Methods, see Supplementary Information).

Numbers of repetitive elements

Analysing 76% of the human genome (using almost all available contigs, Table 1), we estimated that around 43% of the human genome is occupied by four major classes of interspersed repetitive element: (1) short interspersed elements (SINES), (2) long interspersed elements (LINEs), (3) elements with long terminal repeats (LTR elements), and (4) DNA transposons. There are more than 4.3 million repetitive elements in the human genome, with *Alu* and LINE1 (L1) being the most frequent. These estimates largely agree with previous ones^{1,2}. As many repetitive elements would have degenerated to the extent that they cannot be detected by the computer program RepeatMasker (<http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker>), more than 50% of the human genome would have come from insertion of repetitive elements.

Repetitive elements in proteins

Contrary to the belief that a repetitive element insertion into a gene is deleterious and unlikely to survive, many translated repetitive elements are found in proteins (Table 2). From the International Protein Index (ref. 3; <http://www.ensembl.org/IPI>), we derived a new database by eliminating 'isoforms' (due to alternative splicing). We set the expected (*E*) value $< 10^{-80}$ in BLASTing (using tBLASTN) the database by itself and deleted all but one copy of the genes whose chromosome locations overlapped by more than 50%. This procedure reduced the number of 'proteins' in the database from 45,112 to 43,195, of which 15,337 are 'known' proteins and 27,858 are predicted proteins (translated from gene predictions). Because of the stringent conditions used, the chance of misidentifying isoforms is

negligible. The new database probably still contains some isoforms because the chromosomal locations of many sequences are unknown and their isoforms cannot be identified.

We then BLASTed each sequence in the new database against a recent release of RepBase (www.girinst.org). Predicted proteins on average contain many more matches to repetitive element fragments than 'known' proteins (Table 2), suggesting many false positives in gene prediction. This is not a serious problem for 'known' proteins, as they have been translated from genes cloned by traditional methods or from 'genes' that have a high similarity to known genes. Surprisingly, many 'known' proteins also contain (truncated) repetitive elements, especially L1 and *Alu*. A closer look suggests that repetitive elements were usually not inserted into the original open reading frames, but became part of a gene because of alternative splicing, which can sometimes extend or truncate the coding region. L1 has on average the highest *E*-scoring matches (Table 2), indicating that L1-mediated gene evolution may be common. In addition, there is evidence that transduction of 3'-flanking sequences (including exons) is common in L1 retrotransposition⁴, so that L1 might have mediated many exon-shuffling events. Therefore, repetitive elements may have been significant in gene evolution and species differentiation.

To reduce the effect of false gene predictions, we deleted from the database 2,615 predicted proteins that had a significant hit ($E < 10^{-4}$) by a repetitive element and did not have a domain structure other than reverse transcriptase or transposase. The 'cleaned' database contains 15,337 'known' proteins and 25,243 predicted proteins (total, 40,580).

Domain sharing and conservation

A domain is a structural or functional unit in a protein. To investigate the frequency of domain sharing, where the same domain appears in different proteins, we obtained a collection of human, fruitfly, nematode and yeast proteins (15,312, 8,896, 9,254 and 3,136 polypeptides) containing at least one domain; we used the InterPro domain database. In each case of nested domains, only the shortest one was included in the final dataset. There are 1,865, 1,218, 1,183 and 973 domain types in human, fruitfly, nematode and yeast, respectively, and the proportions of mosaic proteins (containing

Table 1 Repetitive elements in the human genome

Type	No. found	Estimated no. in genome	Per cent of genome
SINE (all)	1,404,300	1,841,000	12.5
<i>Alu</i>	1,010,400	1,324,600	10.7
LINE (all)	1,045,800	1,371,100	18.9
L1	661,000	866,600	15.4
DNA	308,800	404,900	2.7
LTR	531,900	697,300	7.9
Other	7,300	9,600	0.1
Total	3,959,200	4,323,900	42.5

The numbers were obtained by using RepeatMasker to mask all contigs assigned to chromosomes. The sequence database used was the 17 July 2000 freeze. Sequencing gaps were removed. Total length of analysed sequences (2,440,850,649bp) is ~76% of the human genome. Per cent of genome means the estimated proportion of the human genome occupied by the repetitive elements under consideration.

Table 2 Repetitive elements in 'known' and predicted proteins

Proteins	$E \leq 10^{-50}$		$10^{-50} < E \leq 10^{-10}$		$10^{-10} < E \leq 10^{-4}$		$10^{-4} < E \leq 10^{-2}$	
	<i>n</i>	Top hit	<i>n</i>	Top hit	<i>n</i>	Top hit	<i>n</i>	Top hit
'Real'	64	Line1	127	<i>Alu</i>	162	<i>Alu</i>	304	Line1
Predicted	870	Line1	1,519	Line1	1,232	<i>Alu</i>	997	<i>Alu</i>
All	1,034	Line1	1,646	Line1	1,394	<i>Alu</i>	1,301	<i>Alu</i>

n, number of cases. 'Known' proteins are translated from genes that have been cloned by traditional methods or have high similarities to known genes. Predicted proteins are translated from genes predicted from genomic sequences using computer programs.

more than one domain type) in the four taxa are 28%, 27%, 21% and 19%.

First, we consider the sharing of domain types (or domain combination), regardless of the order or the number of times a domain appears in a protein; for example, a protein with A-A-B-B-A contains only two domain types and has the combination AB. In our database, the largest number of domain types per protein is nine in human and fruitfly, and seven in nematode and yeast. The frequency of domain sharing is very high among human proteins (Table 3); for example, there are 88 cases where three proteins share two types of domain. There are also many human proteins that share more than one type of domain with *Drosophila*, (slightly less frequently) with *C. elegans*, and (much less frequently) with yeast proteins. But there are only three cases where a combination of more than three domain types is shared by human and yeast proteins and only two of these cases are shared by the four taxa. One of these two cases has a combination of seven domain types; it occurs once in human, nematode and yeast but twice in the fruitfly. It is a carbamoyl-phosphate synthase (EC 6.3.5.5) involved in the first three steps of *de novo* pyrimidine nucleotide biosynthesis (SwissProt accession nos P07259, Q18990, Q9VXD5, P27708).

We now consider the conservation of domain arrangements (the number and order of domains within a protein). There are 3,433, 1,702, 1,248 and 470 distinct arrangements of two or more domains in human, fruitfly, nematode and yeast proteins, respectively. Some

proteins exhibit extensive domain repetition: in human, the largest number of domain types in a protein is nine, but the largest total number of domains in a protein is 130. Many human proteins have identical arrangements (Table 3). In the case of two domain types, many of the human arrangements are shared by fruitfly, (less frequently) by nematode, and (even less frequently) by yeast. The largest domain arrangement shared by all four taxa contains 11 domains, with only two domain types. It is ornithine decarboxylase, which catalyses a rate-limiting step in the biosynthesis of polyamines (SwissProt: P08432, Q94278, Q9V352, P11926). The shared arrangement that has the largest number of domain types (four) contains five domains; it is a sulphonylurea receptor (SURx) in fruitfly, which is a subunit of the ATP-sensitive potassium channel (SwissProt: P53049, Q9U6Z2, Q9V352).

Duplicate genes

Two genes that were derived from a gene duplication are said to be paralogous; two genes (in two species) are orthologous if they were derived from the same gene through speciation. Predicting whether two proteins are paralogous is relatively simple when their sequence identity (*I*) is high (>40% for long sequences) but becomes difficult when *I* is in the medium range (20–35%) or lower, especially for short sequences.

Rost⁵ proposed an empirical formula for clustering proteins in a database (Table 4). Two proteins are assumed to be paralogous if the proportion (*p*) of identical residues over the *L* aligned amino-acid residues between the two proteins is higher than the cut-off point (*p*[†]) defined by the formula. The cut-off point increases as *L* decreases because two unrelated short sequences may by chance have a high *p* value. A common practice in clustering proteins into groups is to use single linkage: if proteins A and B have a *p* higher than *p*[†] and so do proteins B and C, then A, B and C are clustered in the same group, even if the *p* value for A and C does not meet the cut. Applying Rost's formula with *n* = 5 (*n* is a factor to raise the cut-off point) to the 'cleaned' protein database, we found that the largest group contained 15,121 members, which is more than one-third of the database and includes various proteins. Even for *n* = 25 the largest group still contained 4,519 members. Such large groups

Table 3 Domain sharing and order conservation within human and between human and other eukaryotes

Human versus	Domain sharing					Identical domain arrangements			
	No. of proteins sharing domains	No. of cases				Total no. of domains in a protein	No. of identical arrangements/no. of human proteins		
		No. of domain types					No of domain types*		
		1	2	3	>3		2	3	>3
Human	2	214	194	73	61	1	-	-	-
	3	147	88	25	18	2	141/556	-	-
	4	123	38	17	5	3	57/208	21/62	-
	5	67	17	5	3	4	53/168	18/63	4/10
	6	56	19	5	0	5	44/173	11/27	5/16
	>6	377	79	20	5	>5	150/605	66/172	34/78
Fly	1	143	129	32	23	1	-	-	-
	2	134	65	14	12	2	119/337	-	-
	3	97	47	11	5	3	35/98	10/18	-
	4	83	19	7	0	4	28/65	10/24	1/1
	5	51	9	2	2	5	25/74	8/17	5/13
	>5	359	65	14	2	>5	58/137	11/19	12/16
Worm	1	136	92	27	9	1	-	-	-
	2	124	56	11	12	2	89/307	-	-
	3	94	38	9	7	3	28/118	10/24	-
	4	84	17	5	2	4	16/39	6/20	0/0
	5	46	8	2	2	5	16/60	3/8	3/6
	>5	355	61	11	1	>5	43/118	8/16	9/13
Yeast	1	135	51	8	2	1	-	-	-
	2	91	27	5	0	2	51/199	-	-
	3	64	18	2	0	3	9/20	4/12	-
	4	58	5	0	0	4	4/7	3/3	0/0
	5	41	3	0	0	5	3/6	1/2	1/1
	>5	260	24	4	1	>5	36/16	1/3	0/0
Fly, worm and yeast	1	75	24	4	1	1	-	-	-
	2	78	16	3	0	2	26/145	-	-
	3	49	12	1	0	3	4/10	1/3	-
	4	48	3	0	0	4	3/5	0/0	0/0
	5	33	2	0	0	5	3/6	0/0	1/1
	>5	249	21	3	1	>5	7/18	0/0	0/0

*Number of unique domain arrangements/number of human proteins in which these arrangements are found. The second number is larger than the first because many proteins may share the same arrangement. For example, in the case 4/10 (bold numbers) there are four unique arrangements of three-domain proteins with two domain types (for example, the arrangement A-B-A has three domains but only two domain types: A and B) that have been conserved among human, fly, worm, and yeast; in human there are ten such proteins.

Table 4 Protein groups inferred from sequence similarities

Group size	<i>I</i> [†] ≥ 50%*	<i>I</i> [†] ≥ 40%†	<i>I</i> [†] ≥ 30%‡
	No. of groups	No. of groups	No. of groups
1	31,515	28,251	25,237
2	2,041	2,343	2,288
3–5	807	1,069	1,298
6–10	104	170	262
11–20	36	57	86
21–50	14	26	38
51	1		2
69		1	1
71			1
104	1		
122			1
124	1		
129	1		
132		1	
133		1	1
139	1		
221		1	
232		1	
265			1
292			1
331			1
358		1	
479			1

Total number of 'proteins': 40,580 (15,337 'known' proteins and 25,243 predicted proteins). All comparisons with *L* ≤ 20 were excluded.

**I*[†] ≥ 50% for *L* > 40 and *I*[†] ≥ *p*[†] for *L* > 40, where *p*[†] is given by Frost's formula⁴ $p^{\dagger} = 0.01n + 4.8L^{(-0.32n1 + \exp(-L/1000))}$. For *n* = 0, *p*[†] = 72%, 41%, 28% and 24% for *L* = 20, 50, 100 and 150, respectively.

†*I*[†] ≥ 40% for *L* > 70 and *I*[†] ≥ *p*[†] for *L* > 70.

‡*I*[†] ≥ 30% for *L* > 150 and *I*[†] ≥ *p*[†] for *L* > 150.

occur probably because nonhomologous proteins may share the same domains (see above).

We propose to use $I' = I \times \text{Min}(n_1/L_1, n_2/L_2)$, where I is the proportion of identical amino acids in the aligned region (including gaps) between the query (sequence 1) and target (sequence 2) sequences obtained by the alignment program FASTA, L_i is the length of sequence i , and n_i is the number of amino acids in the aligned region in sequence i . The factor $\text{Min}(n_1/L_1, n_2/L_2)$, which means the smaller of n_1/L_1 and n_2/L_2 , takes care of the situation where a high I value is obtained when a short protein shares one or more domains with a longer protein. Another difference between I' and p' is that I' imposes a gap penalty in the aligned region. For short proteins, however, I' may become high by chance and so we impose $I' \geq p'$ with $n = 5$.

Table 4 shows the protein groups inferred from our formula. $I' \geq 50\%$ corresponds to Dayhoff's definition of protein families. The largest group (139 members) contains the L1 reverse transcriptase (RT) and sequences with high I' values with L1 RT. This is surprising, but many 'known' and predicted proteins contain (truncated) L1 RT; note also that many L1 RTs may still be nearly intact in the human genome. The second largest group (129 members) contains 91 immunoglobulin heavy chains, 1 rheumatoid factor, 6 unnamed proteins and 31 predicted proteins; the third (124 members) contains 85 immunoglobulin light chains, 2 heavy chains, 1 microfibrillar protein, 2 unnamed proteins and 34 predicted proteins; the fourth (104 members) contains 38 zinc finger proteins, 6 unnamed proteins and 60 predicted proteins; and the fifth (51 members) contains 16 olfactory receptors and 35 predicted proteins. This criterion identifies 3,007 families, 2,041 of which are two-protein families. These should be taken as minimum estimates because many human genes remain unidentified. For $I' \geq 40\%$, the zinc finger group becomes the largest, and the L1 RT and olfactory receptor groups become the second and third largest. For $I' \geq 30\%$, the five largest groups are zinc finger proteins, olfactory receptors, immunoglobulins (both light and heavy chains), L1 RTs and keratins. For $I' \geq 25\%$, some of the largest groups become very heterogeneous, indicating that at this level of similarity it requires a more rigorous analysis to determine whether two proteins are related.

The $I' \geq 30\%$ criterion identifies 3,982 superfamilies (Table 4).

Although some of the groupings may be false positives, this number may represent a minimum estimate because many human genes remain unidentified and because many of the proteins in the 'singleton' groups (25,237) may actually be related to each other. Taking the data at face value, the proportion of 'singleton' groups is $25,237/40,580 = 62\%$ of the total 'proteins' in our 'cleaned' database. This may be an overestimate, but should be taken cautiously because many of the 'singletons' may be false positive and because the total number of human genes remains unknown.

Our analysis has provided some insights into the evolutionary genomics of the human genome. There are many repetitive elements in our genome (Table 1), and they may have been very important in the evolution of mammalian proteins (Table 2). Domain sharing is common among proteins, and many domain arrangements have been conserved (Table 3). But many challenges remain. For example, as the number of human genes is still unknown, it remains unclear how many human genes exist as single copies. Reliable annotation of the human genome and clean databases of human genes and proteins are required for a rigorous analysis. In addition, better tools are needed for analysis. Single linkage does not seem appropriate for clustering proteins. Finally, better methods are needed for deciding whether two proteins are homologous, especially for short proteins. □

1. Smit, A. F. A. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**, 657–663 (1999).
2. Gu, Z., Wang, H., Nekrutenko, A. & Li, W.-H. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequences. *Gene* **259**, 81–88 (2000).
3. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
4. Goodier, J. L., Ostertag, E. M., Kazazian, H. H. Jr Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**, 653–657 (2000).
5. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94 (1999).

Supplementary information is available from Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank R. Stevens for letting us use Argonne computers, E. Birney for help and NIH for research support.

Correspondence should be addressed to W.-H.L. (e-mail: whli@uchicago.edu).

Cancer and genomics

P. Andrew Futreal*, Arek Kasprzyk†, Ewan Birney†, James C. Mullikin‡, Richard Wooster*§ & Michael R. Stratton*§

* Cancer Genome Project and ‡ Informatics Division, Sanger Centre, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK

† EBI-EMBL, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

§ Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK

Identification of the genes that cause oncogenesis is a central aim of cancer research. We searched the proteins predicted from the draft human genome sequence for paralogues of known tumour suppressor genes, but no novel genes were identified. We then assessed whether it was possible to search directly for oncogenic sequence changes in cancer cells by comparing cancer genome sequences against the draft genome. Apparently chimaeric transcripts (from oncogenic fusion genes generated by chromosomal translocations, the ends of which mapped to different genomic locations) were detected to the same degree in both normal and neoplastic tissues, indicating a significant level of false positives. Our experiment underscores the limited amount and variable quality of DNA sequence from cancer cells that is currently available.

All cancers are caused by abnormalities in DNA sequence. Throughout life, the DNA in human cells is exposed to mutagens and suffers mistakes in replication, resulting in progressive, subtle changes in the DNA sequence in each cell. Occasionally, one of these somatic mutations alters the function of a critical gene, providing a growth advantage to the cell in which it has occurred and resulting in the emergence of an expanded clone derived from this cell. Additional mutations in the relevant target genes, and consequent waves of clonal expansion, produce cells that invade surrounding tissues and metastasize. Cancer is the most common genetic disease: one in three people in the western world develop cancer, and one in five die from it¹.

Around 30 recessive oncogenes (tumour suppressor genes) and more than 100 dominant oncogenes have been identified. In the past, the most successful way to identify such genes was to narrow their location to a small part of the genome using mapping strategies, and then to screen candidate genes in the region for mutations in cancer cases. However, this strategy has its limitations. Mapping information can be confusing or misleading. Moreover, some cancer genes leave no obvious 'identifiers' in the genome and therefore cannot be readily positioned using such maps (for example, dominant oncogenes that are activated by single nucleotide substitutions leading to a single amino-acid change). These mutated genes are essentially invisible to conventional detection and their identification has usually depended upon selection of likely candidates on the basis of the biological features associated with cancer.

How will the genome sequence help us to identify the remaining cancer genes? One possibility is to generate a longer list of plausible candidates by searching for paralogues of known cancer genes. To this end we searched a protein set which represents the current 'best guess' of proteins encoded by the human genome² for paralogues of known recessive oncogenes/tumour suppressor genes (see Supplementary Information; Table 1). Although we detected most of the known family members, we found no convincing evidence of new ones at the level of stringency of this search (50% or greater amino-acid identity over a minimum of 50 amino acids). The lack of detectable paralogues could be due to deficiencies of the protein set, despite the fact that new paralogues have been found for many other families². To confirm our results, we ran an exhaustive search of one of the genes, *TP53*, against the total draft genome sequence, considering every possible gene prediction and reading frame³. This additional search revealed no unknown paralogues.

The lack of novel paralogues may reflect the biological and medical importance of these gene families; most of their members may have already been found. But additional paralogues could be

hiding in unsequenced regions of the genome or may have gone undetected owing to the deficiencies of gene prediction. Clearly, detection of paralogues also depends upon the search criteria used. Predictably, when we lowered the stringency of the search the number of hits increased substantially—but many of these have questionable biological validity.

Was this really a useful way of generating likely candidate cancer genes? Mutated cancer genes do recur in certain gene families (for example, signalling kinases and GDP binding proteins), but these families are often large and, in most cases, only a minority of members are implicated in cancer. In fact, the diversity of structure and function of cancer genes is striking. For example, hardly any of the known recessive oncogenes have strong homology to any other, and their proteins are associated with diverse biological and biochemical functions. Moreover, many close relatives of important cancer-related genes (for example, the genes for p73 and p63, which show sequence similarity to *TP53* (ref. 4)) are not known to be mutated in cancer (although it is possible that they are significantly altered by changes in expression). So we may learn more about the mutations driving cancer if we are not too heavily influenced by past experience. Instead, we should persevere in exploring every gene or protein, whatever its structure or putative function, as a possible candidate.

By simply mining the genome sequence for similarity to known cancer genes, we may miss an opportunity. In addition to many more gene sequences, the working draft contains information concerning the organization of genes, their structures and ordering along the chromosomes. Cancers are characterized by disruption and disorganization of genomes. Perhaps we could use the genome sequence as a template against which we can detect structural alterations of the genome in cancer cells.

To do this we need to compare the working draft (and ultimately the finished sequence) with corresponding sequences from cancer cell genomes. However, there is very little cancer genome sequence available, and what is available is patchy. The largest body of sequence from cancer genomes originates from programs that sample clones in complementary DNA libraries constructed from neoplastic tissues (for example, the Cancer Genome Anatomy Project (CGAP, <http://www.ncbi.nlm.nih.gov/CGAP/>)). In principle, we could try and compare these with the working draft for the somatic base substitutions and small insertions and deletions that often result in inactivation of tumour suppressor genes or activation of dominantly acting oncogenes. In practice, however, these databases contain relatively little sequence and do not sample most transcripts in any single tumour. Moreover, the available sequence is mostly from untranslated regions of genes (whereas the cancer-causing mutations cluster in the coding regions). Even if the rare,

meaningful somatic mutations could be detected, they would be buried in the debris of sequence errors (both in the cDNA libraries and in the genomic sequence) or camouflaged in a forest of innocuous polymorphisms.

We attempted to use these cDNA library sequences in conjunction with the working draft to look for a different type of alteration in cancer. Gene rearrangements that result in activation of oncogenes can arise as a result of chromosomal translocation. This type of abnormality is common in leukaemias, lymphomas and sarcomas^{5,6}, and often results in the formation of a chimaeric transcript, the product of a fusion gene derived from portions of transcribed genes on either side of the chromosomal breakpoints (although in some instances the translocation simply results in dysregulation of an intact transcript). Intriguingly, this pattern of oncogenesis has not been frequently documented amongst most of the common, adult epithelial cancers. Whether the rearrangements exist but are hidden in the disorganized complexity of epithelial cancer cell genomes or are simply not present in epithelial cancers is a question that may be addressed in the near future.

To look for such gene rearrangements we obtained all the sequences from the CGAP program (derived from cDNA libraries constructed from both normal tissue samples and neoplastic samples) and selected those clones from which two sequences had

been obtained (normal, 215,889 clones; cancer, 25,446 clones). Most of these sequences were from either end of the clones. We then looked in the genomic sequence for matches to these paired cDNA sequences that would allow their chromosomal localization (see Supplementary Information).

Most pairs of sequences from the same cDNA clone mapped to the same position in the genome, as would be expected if they had originated from a single normal transcript. But there were a few pairs derived from a single cDNA clone that mapped to two different parts of the genome. Could they represent transcripts of chimaeric genes generated by chromosomal translocation? Possibly, but this strategy has limitations. Even this conceptually simple experiment was dogged by the intrinsic complexities of the genome, such as low-frequency repeats and multiple, high-fidelity copies of some genes. Moreover, our analyses yielded proportionally the same number of apparently chimaeric transcripts derived from normal tissues as from cancers (3% of starting clones in both cases), indicating a significant rate of false positives. These could result from chimaeric clones arising as artefacts of cDNA library construction, mistracking of sequencing gels, errors in annotating and curating databases or misassembly of the draft sequence.

Of course, for this experiment we used a resource that had not

Table 1 Recessive oncogenes used in the search for paralogues

Symbol	Accession	Name	Paralogues identified*	Cancer syndrome	Cancer types (germline and/or somatic mutations)
<i>APC</i>	P25054	Adenomatous polyposis of the colon gene	APC-like gene	Familial polyposis of the colon	Colorectal, pancreatic cancers, desmoids, hepatoblastoma
<i>BRCA1</i>	P38398	Familial breast/ovarian cancer gene 1	None	Hereditary breast/ovarian cancer	Hereditary breast/ovarian cancers
<i>BRCA2</i>	P51587	Familial breast/ovarian cancer gene 2	None	Hereditary breast/ovarian cancer	Hereditary breast/ovarian cancers
<i>CDH1</i>	P12830	Cadherin 1 gene	N-cadherin, P-cadherin, R-cadherin	Familial gastric carcinoma	Lobular breast cancer
<i>CDKN2A</i>	P42771	Cyclin-dependent kinase inhibitor 2A (p16) gene	None	Cutaneous malignant melanoma 2	Melanoma, other tumour types
<i>CYLD</i>	CAB93533	Familial cylindromatosis gene	None	Familial cylindromatosis	Cylindromas
<i>EP300</i>	Q09472	300-KD E1A-binding protein gene	Rubinstein-Taybi gene	n/a	Colorectal, breast, pancreatic cancers
<i>EXT1</i>	Q16394	Multiple exostoses type 1 gene	EXT-like 1	Multiple exostoses type 1	Exostoses, osteosarcoma
<i>EXT2</i>	Q93063	Multiple exostoses type 2 gene	EXT-like 3	Multiple exostoses type 2	Exostoses, osteosarcoma
<i>CDKN1C</i>	P49918	Cyclin-dependent kinase inhibitor 1C gene	None	Beckwith-Wiedemann syndrome	Wilms' tumour, rhabdomyosarcoma
<i>STK11</i>	Q15831	Serine/threonine kinase 11 gene	None	Peutz-Jeghers syndrome	Jejunal hamartomas, ovarian tumours, testicular and pancreatic cancers
<i>MAP2K4</i>	P45985	Mitogen-activated protein kinase kinase 4	Multiple map kinase family members	n/a	Pancreatic, breast, colon cancers
<i>MEN1</i>	O00255	Multiple endocrine neoplasia type 1 gene	None	Multiple endocrine neoplasia type 1	Parathyroid/pituitary adenoma, islet cell carcinoma, carcinoid tumours
<i>MLH1</i>	P40692	<i>E. coli</i> MutL homologue gene	None	Familial non-polyposis colorectal cancer	Colorectal, endometrial, ovarian cancer
<i>MSH2</i>	P43246	<i>E. coli</i> MutS homologue 2 gene	None	Familial non-polyposis colorectal cancer	Colorectal, endometrial, ovarian cancer
<i>NF1</i>	P21359	Neurofibromatosis type 1 gene	None	Neurofibromatosis type 1	Neurofibroma, glioma
<i>NF2</i>	P35240	Neurofibromatosis type 2 gene	Moesin	Neurofibromatosis type 2	Meningioma, acoustic neuroma
<i>PRKAR1A</i>	P10644	Protein kinase A type 1- α regulatory subunit gene	None	Carney complex	Myxoma, endocrine tumours
<i>PTCH</i>	Q13635	Homologue of <i>Drosophila patched</i> gene	Patched 2	Nevoid basal cell carcinoma syndrome	Basal cell carcinoma, medulloblastoma
<i>PTEN</i>	O00633	Phosphatase and tensin homologue gene	None	Cowdens syndrome	Hamartomas, glioma, prostate and endometrial cancers
<i>RB1</i>	P06400	Retinoblastoma gene	None	Familial retinoblastoma	Retinoblastoma, sarcomas, breast cancer, small cell lung cancer
<i>SDHD</i>	O14521	Succinate dehydrogenase cytochrome B small subunit gene	None	Familial paraganglioma	Paraganglioma
<i>MADH4</i>	NP_005350	Homologue of <i>Drosophila mothers against decapentaplegic 4</i> gene	None	Juvenile polyposis	Gastrointestinal polyps, colorectal and pancreatic cancers
<i>SMARCB1</i>	Q12824	Swi/Snf5 matrix-associated actin-dependent regulator of chromatin gene	None	Rhabdoid predisposition syndrome	Malignant rhabdoid tumours
<i>TP53</i>	P04637	Tumour suppressor p53 gene	p73, p63 genes	Li-Fraumeni syndrome	Sarcoma, adrenocortical carcinoma, glioma, other tumour types
<i>TSC1</i>	Q92574	Tuberous sclerosis 1 gene	None	Tuberous sclerosis 1	Hamartomas, renal cell carcinoma
<i>TSC2</i>	P49815	Tuberous sclerosis 2 gene	None	Tuberous sclerosis 2	Hamartomas, renal cell carcinoma
<i>VHL</i>	P40337	Von Hippel-Lindau syndrome gene	None	Von Hippel-Lindau syndrome	Renal cell carcinoma, hemangioma, pheochromocytoma
<i>WT1</i>	P19544	Wilms tumour 1 gene	Zinc finger-containing genes	Familial Wilms tumour	Wilms tumour

* 50% identity over at least 50 amino acids.

been designed to support such investigations, so it is not surprising that it did not bear much fruit. In addition to the problems of false positives, relatively few clones have been sequenced from most of the libraries, many of the libraries are not normalized (leading to undersampling of less abundant transcripts) and many of the sampled cDNAs are not full length. This was illustrated by the fact that when we searched the CGAP annotations of libraries constructed from five cancer types with known chimaeric genes, we found only one of the ten genes involved in the five chimaeric transcripts. This exercise underscores the point that elucidating the complexity of cancer at the genomic level will require much more sequence data from cancer genomes, which will need to be configured appropriately for the task at hand.

So, the working draft will not immediately reveal the natures of the abnormalities in cancer cell genomes. To facilitate these analyses we will need the finished sequence, which will form a structural framework for a new generation of massive-scale comparisons of cancer cell and normal genomes. Ultimately, it may also prompt a shift away from strategies that depend upon primary genomic localization and allow systematic genome-wide searches for mutations. New technology will be required; there is no single technology at present that will detect all the types of abnormality

(large deletions, rearrangements, base substitutions, small insertions and deletions, amplifications, and epigenetic changes such as methylation) that are present in cancer cells. Sequencing of genomic libraries constructed from cancer genomes would come closest to this goal, but given the diversity of cancers and the effort and cost required to obtain reasonable coverage of a human genome this is a daunting challenge. □

1. Higginson, J., Muir, C. & Munoz, N. *Human Cancer: Epidemiology and Environmental Causes* (Cambridge Monographs on Cancer Research, Cambridge, UK, 1992).
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–548 (2000).
4. Kaelin, W. G. Jr The p53 gene family. *Oncogene* **18**, 7701–7705 (1999).
5. Rowley, J. D. The critical role of chromosome translocations in human leukemias. *Annu. Rev. Genet.* **32**, 495–519 (1998).
6. Bell, R. S., Wunder, J. & Andrulis, I. Molecular alterations in bone and soft-tissue sarcoma. *Can. J. Surg.* **42**, 259–266 (1999).

Supplementary information is available on *Nature's* World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of *Nature*.

Correspondence and requests for materials should be addressed to M.R.S. (e-mail: mrs@sanger.ac.uk).

Human disease genes

Gerardo Jimenez-Sanchez*, Barton Childs* & David Valle*†

* Department of Pediatrics, McKusick-Nathans Institute of Genetic Medicine, and † Howard Hughes Medical Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

The complete human genome sequence will facilitate the identification of all genes that contribute to disease. We propose that the functional classification of disease genes and their products will reveal general principles of human disease. We have determined functional categories for nearly 1,000 documented disease genes, and found striking correlations between the function of the gene product and features of disease, such as age of onset and mode of inheritance. As knowledge of disease genes grows, including those contributing to complex traits, more sophisticated analyses will be possible; their results will yield a deeper understanding of disease and an enhanced integration of medicine with biology.

To test the proposal that classifying disease genes and their products according to function will provide general insight into disease processes^{1,2}, we have compiled and classified a list of disease genes. To assemble the list, we began with 269 genes identified in a survey of the 7th edition of *Metabolic and Molecular Bases of Inherited Disease*². We then searched the 'morbidity map' and allelic variants listed in the *Online Mendelian Inheritance in Man*³ (OMIM), an online resource documenting human diseases and their associated genes

(www.ncbi.nlm.nih.gov), and increased the total disease gene set to 923. This sample included genes that cause monogenic disease (97% of the sample) and genes that increase susceptibility for complex traits. We excluded genes associated only with somatic genetic disease (such as non-inherited forms of cancer) or the mitochondrial genome.

Functional classification

We categorized each disease gene according to the function of its

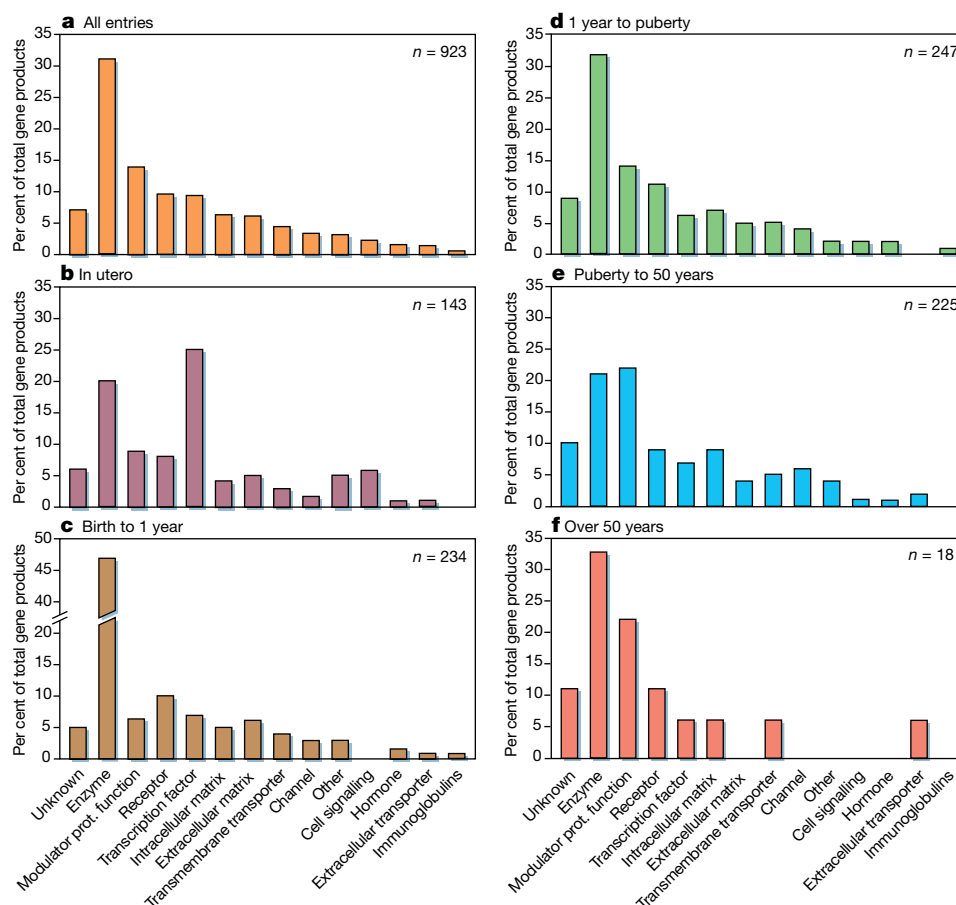


Figure 1 The functions of the protein products of disease genes. **a**, The entire disease gene set. **b–f**, Disease genes stratified according to the typical age of onset of the disease phenotype. The fraction of disease genes encoding transcription factors in the *in utero* onset disorders (25%) differs from the fraction encoding transcription factors for disorders with onset after birth (6%; $\chi^2 = 49.4$, $P < 0.001$). Similarly, the fraction of disease genes encoding enzymes causing a disorder with onset in the first year of life (47%) is different from the fraction encoding enzymes causing disorders with other ages of onset (25.8%; $\chi^2 = 35.8$, $P < 0.001$).

protein product (see Supplementary Information). Our approach differed in two ways from that used by the International Human Genome Sequencing Consortium (IHGSC) to annotate the working draft human sequence⁴. First, we focused on the function of the protein itself without consideration of its biological context, whereas the IHGSC used the classification employed by the Gene Ontology project⁵ which integrates three aspects of function: biochemical activity, biological process and subcellular location. Second, our functional designations were largely informed by features of pathology, whereas those of the IHGSC were almost entirely based on sequence homology to proteins of known function in model organisms. We also scored each disease gene for features related to clinical presentation, including age of onset, mode of inheritance, frequency, severity, extent of tissue involvement and association with malformations.

The results of our functional classification of the proteins encoded by 923 disease genes are shown in Fig. 1a. The largest functional category, comprising genes encoding enzymes, accounts

for 31.2% of the total. This represents about twice as many as the next highest category, designated modulators of protein function (13.6 %), which includes proteins that stabilize, activate, fold or otherwise influence the function of a second protein. Each of the remaining 12 categories accounts for less than 10% of the total sample. The abundance of enzymes in the disease gene set may reflect some historical bias towards metabolic disorders in the study of human inherited disease². In contrast, only 15% of 114 positionally cloned genes (updated from <http://genome.nhgri.nih.gov/clone/>) encode enzymes, but this set may have its own biases (see Supplementary Information). Indeed, protein domains associated with enzymes were identified in 27% of 8,360 *Drosophila* proteins scored for these motifs⁶. This observation suggests that in higher eukaryotes the fraction of genes encoding enzymes may be 25–30%, or close to the fraction identified in our disease gene set.

Gene function and disease characteristics

We analysed the disease gene set for evidence of correlations

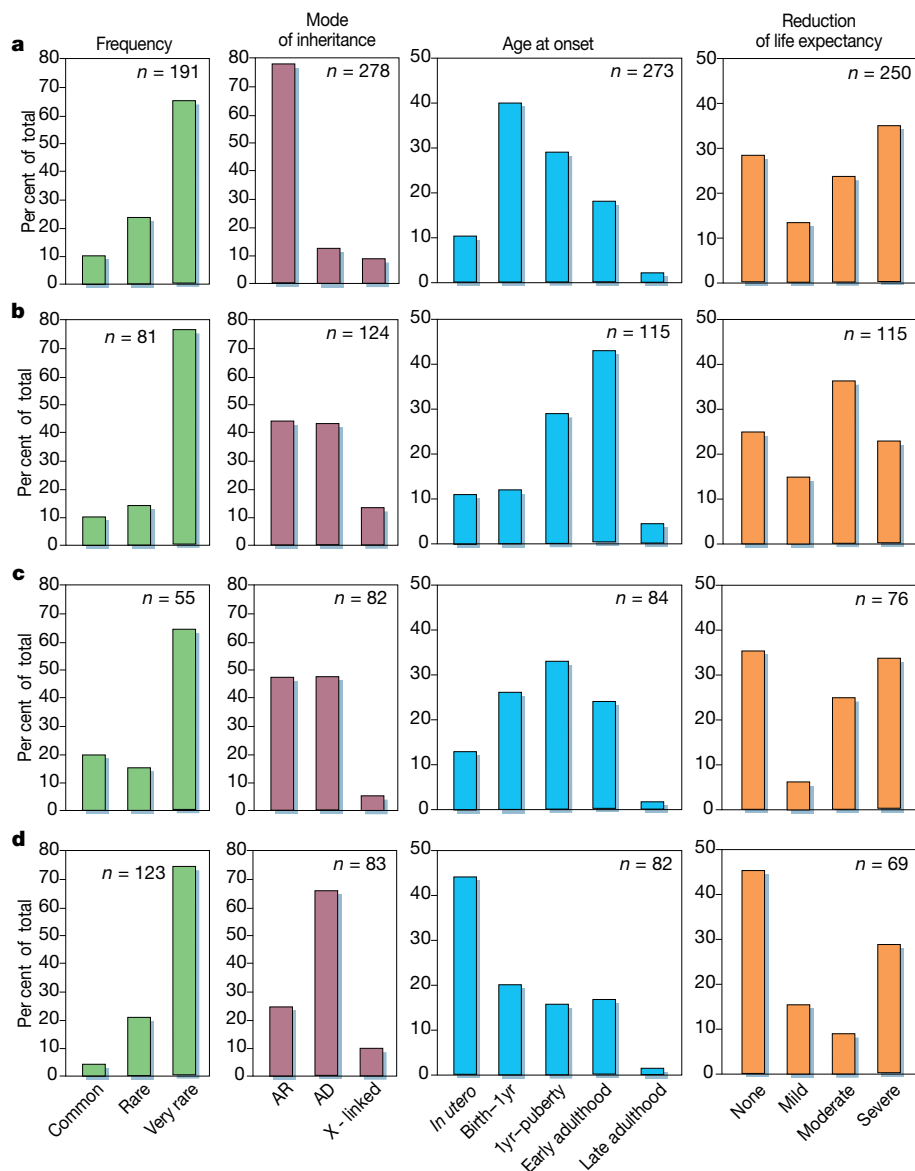


Figure 2 Characteristics of disease arranged by function of the protein encoded by the disease gene. **a**, Disease genes encoding enzymes; **b**, disease genes encoding modifiers of protein function; **c**, disease genes encoding receptors; **d**, disease genes encoding transcription factors. The columns of disease features are labelled at the top. AR, autosomal recessive; AD, autosomal dominant; early adulthood, puberty to <50 years; late adulthood, >50 years.

between the function of a gene product and the age of onset of its associated disease (Fig. 1). Several aspects of this analysis are of interest. First, diseases associated with genes encoding proteins in all the functional categories can appear at any stage of life. The only apparent exception is for diseases presenting after 50 years of age (Fig. 1f) but the sample of genes in this category is small and a more general distribution of protein function may emerge as the number increases. Second, genes encoding transcription factors are over-represented among genes causing genetic disease with onset *in utero* (Fig. 1b). This concentration of diseases resulting from abnormalities of transcription factors probably reflects the important role of these proteins in orchestrating development. It is therefore not surprising that genes encoding transcription factors account for more than 30% of the genes associated with malformation phenotypes (see Supplementary Information).

An extraordinarily high fraction of diseases with onset in the first year of life are caused by defects in genes encoding enzymes (47%; Fig. 1c). This too fits with biological expectations and clinical evidence. The developing fetus has access to its mother's metabolic homeostatic systems through the placenta. Thus, infants with inborn errors caused by enzyme deficiencies are typically normal at birth and develop symptoms only after the defect in their homeostatic system is exposed by demands on their own metabolism⁷. The fraction of disease genes encoding enzymes falls with later disease onset (Fig. 1d–f). Disorders with onset after age 50 are an apparent exception, with the fraction of genes encoding enzymes increasing to more than 33%. But the number of disorders in this category (18) is small, and our understanding of the genes that contribute to complex traits with onset in this age range is limited. Three of the six genes encoding enzymes in this age of onset category are variants identified as susceptibility alleles rather than true disease-producing alleles.

We divided the disease gene list by function and compared disease characteristics including frequency, mode of inheritance, age at onset and reduction in life expectancy. Figure 2 shows the results for the four largest functional categories. Regardless of category, diseases caused by most genes in this analysis are rare or very rare in frequency. In part, this reflects our preliminary knowledge of the genes that contribute to common complex traits. Comparison of the inheritance patterns shows that disorders caused by genes encoding enzymes are primarily recessive, whereas those caused by genes encoding modifiers of protein function and receptors are split more-or-less evenly between recessives and dominants. Disorders caused by genes encoding transcription factors, by contrast, are more likely to be dominant. These results fit well with our understanding of how proteins of various functions contribute to development and homeostasis.

Interestingly, each of the four functional categories has a different peak age at onset. For transcription factors the peak is *in utero*; for

enzymes it is in year 1; for receptors it is between year 1 and puberty; and for modifiers of protein function it is in early adulthood. These correlations provide biological support for the validity of the functional characterization and they hint at additional principles of disease. Perhaps disorders of receptors are most likely to present in childhood because this is a time of rapid growth and, especially during puberty, of intense signalling activity between various cells and tissues. Similarly, disorders involving modifiers of protein function may present later in life because the homeostatic systems are not completely disrupted by these defects; rather, they respond in ways that are less congruent with the demands placed on the organism and so become symptomatic more gradually. Finally, there is no apparent relationship between functional category and reduction in life expectancy. This may reflect a true lack of correlation or it may indicate that larger numbers and more sophisticated characterization of disease severity are required to discern such relationships.

Better functional annotation of the human genome and a comprehensive list of human disease genes should lead to much greater integration of medicine and biology. We believe that increasing knowledge of the genes associated with diseases will allow researchers to address more complicated issues, including the relative contributions to disease of genes in the core biological set shared by all species and those encoding proteins specific to humans⁸; how sequence features (such as conservation and polymorphism) relate to disease characteristics; and how protein function relates to the outcome of clinical treatment⁹. □

1. Childs, B. & Valle, D. Genetics, biology and disease. *Annu. Rev. Genomics Hum. Genet.* **1**, 1–19 (2000).
2. Jimenez-Sanchez, G., Childs, B. & Valle, D. in *The Metabolic and Molecular Bases of Inherited Disease* (eds Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D.) 167–174 (McGraw-Hill, New York, 2001).
3. Antonarakis, S. E. & McKusick, V. A. OMIM passes the 1,000 disease gene mark. *Nature Genet.* **25**, 11 (2000).
4. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
5. The Gene Ontology Consortium. Gene ontology: Tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
6. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
7. Brusilow, S. W., Valle, D. & Arn, P. H. in *Current Therapy in Neonatal Perinatal Medicine* (ed. Nelson, N. M.) 164–169 (B. C. Decker, Philadelphia, 1989).
8. Chervitz, S. A. *et al.* Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* **282**, 2022–2028 (1998).
9. Treacy, E., Childs, B. & Scriver, C. R. Response to treatment in hereditary metabolic disease: 1993 survey and 10-year comparison. *Am. J. Hum. Genet.* **56**, 359–367 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank J. Amberger for assistance with OMIM, J. A. Escamilla for statistical advice and S. Muscelli for preparation of the manuscript. D.V. is an Investigator in the Howard Hughes Medical Institute.

Correspondence should be addressed to D.V. (e-mail: dvalle@jhmi.edu).

Computational comparison of two draft sequences of the human genome

John Aach^{*†}, Martha L. Bulyk[‡], George M. Church[†], Jason Comander[§], Adnan Derti^{†||} & Jay Shendure^{*†}

[†] The Lipper Center for Computational Genetics, [‡] Program in Biophysics, [†] Department of Genetics and [§] Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA

^{||} Program in Bioinformatics, Boston University, Boston, Massachusetts 02215, USA

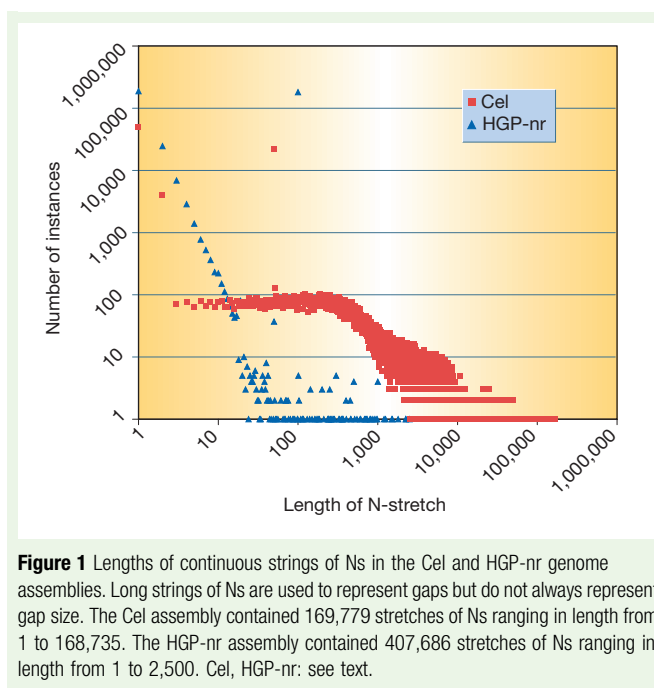
^{*} These authors contributed equally to this work

We are in the enviable position of having two distinct drafts of the human genome sequence. Although gaps, errors, redundancy and incomplete annotation mean that individually each falls short of the ideal, many of these problems can be assessed by comparison. Here we present some comparative analyses of these drafts. We look at a number of features of the sequences, including sequence gaps, continuity, consistency between the two sequences and patterns of DNA-binding protein motifs.

The two draft sequences of the human genome were generated by the Human Genome Project (HGP)¹ and Celera Genomics². Genome sequencing entails direct sequencing of DNA fragments and assembly of the fragment sequences into larger units on the basis of their overlaps (shotgun assembly). The HGP used a hierarchical mapping and sequencing approach, involving generation of a series of overlapping clones that cover the entire genome and shotgun sequencing of each clone. The genome sequence was reconstructed by assembling the fragments on the basis of sequence overlap and mapping and chromosomal position information on the clones. Celera Genomics used a whole-genome shotgun sequencing approach, without generating a series of overlapping clones, but also incorporated HGP information where available.

We worked with three versions of genome sequence. (1) Data from the individual large-insert clones sequenced by the publicly sponsored HGP is available from the National Center for Biotechnology Information (NCBI; <http://www.ncbi.nlm.nih.gov>), and is denoted here as HGP-all. HGP-all comprises 4.8 gigabases (Gb) from 34,084 large-insert clones (identified as phase 0 (raw) to phase 3 (finished) sequence) and is highly redundant because it contains sequences from a collection of overlapping clones, as well as other sources whose overlaps have not been fully resolved through assembly. (2) Several genome-wide assemblies have been produced by merging the data in HGP-all. We studied an assembly comprising 2.9 Gb in 6,094 sequences—designated here as HGP-nr ('non-redundant')—in which clearly identifiable redundancies were eliminated. Several other genome assemblies have also been generated; these assemblies are not analysed here. (3) We used Celera Genomics' 'Human Genome D'², which represents 2.9 Gb in 54,061 sequences, denoted here as Cel (which we obtained through an academic licence to the Celera database). We did not analyse their larger unassembled databases of 23.1 million human fragments and 2.8 million polymorphic variants. NCBI has made rapid progress towards annotating HGP-nr for genes on the basis of their RefSeq database³, which is a manually curated collection of mRNA sequences from known genes, but at the time of writing (December 2000) the Cel sequence had no directly linked annotation.

Cel and HGP-nr use long strings of Ns to indicate gaps in assemblies (Fig. 1). Although the two draft genome sequences are similar in size, HGP-nr contains fewer unidentified bases than Cel (0.65% versus 8.7%, respectively; the 1.0-Gb HGP-all phase 3 sequence contains even fewer, 0.005%), largely because gaps are annotated differently in the two sequences. Thus, when one removes the unidentified bases (Ns), the amount of specified nucleotide sequence is 2.84 Gb for HGP-nr and 2.66 Gb for Cel. HGP-nr



contains 181,079 strings of 100 Ns to represent gaps but contains strings of up to 2,500 Ns; Cel contains 21,684 strings of 50 Ns but contains strings of up to 168,735 Ns. Figure 2 shows the sizes of continuous segments of sequence not broken by strings of 20 or more Ns (ungapped sequences). Whereas Cel uses the lengths of very long strings of Ns to represent the estimated sizes of large gaps, HGP-nr generally simply uses strings of 100 Ns and does not use string length to represent gap size. Therefore the smaller continuous strings of Ns in HGP-nr does not imply that gaps in HGP-nr are smaller than in Cel.

As another indicator of the continuity of the assemblies, we examined the ten genes in the RefSeq database with the largest messenger RNAs and tested whether their coding sequences could be found on single contigs, the largest continuous sections of sequence (possibly interrupted by N strings) generated by sequence assembly (Fig. 2). We used the BLAST search algorithm⁴ to locate the best contig matches for the first and last 500 base pairs (bp) of the coding regions of these ten genes. Six of these genes in HGP-nr, and seven in Cel, had both ends on the same contig. In HGP-nr, *ACF7* and *RYR2* had ends in different contigs, while *NEB* and *MUC2*

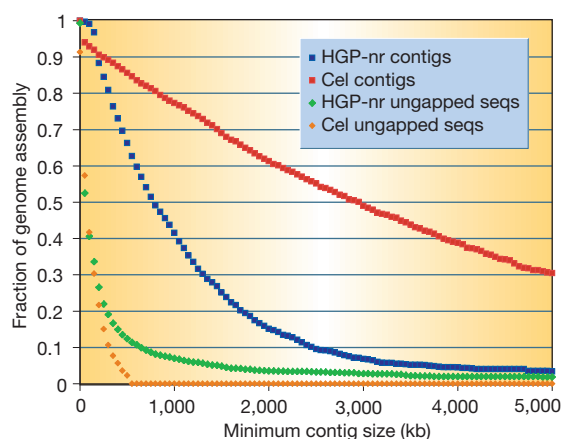


Figure 2 Cumulative histogram of the fraction of assembled genome sequence present in contigs and in ungapped continuous sequences that exceed a given length in the HGP-nr and Cel assemblies. Contigs are the largest continuous segments of sequence generated by assemblies. The Cel assembly consisted of 2.9 Gb in 54,061 contigs. The HGP-nr assembly consisted of 2.9 Gb in 6,094 contigs. Contigs may contain gaps represented by strings of Ns (Fig. 1). Here ungapped continuous sequences (ungapped seqs) are counted as maximal stretches of sequence that do not contain strings of 20 or more Ns. Cel, HGP-nr: see text.

matched a contig for one end but a matching sequence could not be located for the other. In Cel, *ACF7* and *MUC2* had ends in different contigs, and a smaller coding region (*TNNT2*) appeared to have one end in a Cel contig and the other in a stretch of Ns and an unfinished fragment. These results indicate that both sequence assemblies have limited ability to build contigs covering the longest stretches of DNA that cells themselves can transcribe continuously.

Comparing unique stretches of sequence

Oligonucleotides that occur only once in the genome can be useful for making specific probes for genomic DNA or messenger RNA, or as the 3' ends of specific primers for polymerase chain reaction (PCR) amplification. They can also provide a high-level statistical view of how much sequence content the two draft genome sequences have in common. We analysed stretches of 15 nucleotides, referred to as 15-mers. In 6 Gb of genomic sequence (with both strands considered), the number of occurrences of an arbitrary 15-mer (of which there are 4^{15} or slightly more than one billion) can be estimated as following a Poisson distribution with mean of about 6 (see Supplementary Information). Disregarding splice junctions, a typical mRNA sequence of 2000-bp should therefore contain about five unique 15-mers and have a 99% likelihood of having at least one. Using a computer algorithm to analyse the two genome assemblies, we considered every possible 15-mer and determined whether it occurred 0, 1 (unique) or multiple times. Sequence and assembly errors, gaps, redundancies, splice junctions and polymorphisms all affect whether a given 15-mer will be counted as unique. We therefore refer to 15-mers seen only once as candidate unique 15-mers (cu-15s). Assuming a combined rate of sequence error and polymorphism of 0.1% per base, we estimate a cu-15 false positive rate of 0.14% and a cu-15 false negative rate of 9.0% (see Supplementary Information). A false positive is a multiply occurring 15-mer that is detected as unique because variations have made all but one occurrence appear different; a false negative is a unique 15-mer that is detected as multiple because of variations in similar 15-mers.

We found 169,609,634 cu-15s in the 2.9-Gb Cel sequence and 160,311,078 in the 2.9-Gb HGP-nr sequence. Of the cu-15s in the two sequences, 19,270,620 are found only in Cel and 17,527,980 are

found only in HGP-nr; therefore, nearly 11% of cu-15s in each sequence are not shared with the other. Because the analysis is expected to have a false negative rate of 9% for each database, this suggests that the true amount of sequence present in one database and not the other is about 0.14% for both HGP-nr and Cel (see Supplementary Information). Cel and HGP-nr therefore contain similar amounts of unique sequence, and most unique sequences are common to both. We also looked for unique 15-mers representing 10,292 mRNA sequences from the RefSeq database³ (see Supplementary Information); we found 2,526,912 when compared against Cel and 2,372,185 when compared against HGP-nr. Again the Cel and HGP-nr results are comparable.

By analysing cu-15s in coding sequences we found additional differences between the Cel and HGP-nr draft genome assemblies. For instance, the coding regions for the genes *LOC63301* and *GR3*, which are 99% identical over their 1,113 bases, are annotated in HGP-nr as being located on two distinct chromosomes: contig NT_008902 (chromosome 10) and contig NT_023464 (chromosome 6), respectively. However, we found that these sequences are contained on only a single Cel contig (GA_x2HTBL4-GEJJ:1..500000) and, consistent with this, we found 396 cu-15s in the Cel database shared by both sequences. Analysis of the HGP-nr sequence using 'BLAST2 sequences' (ref. 5) indicates that the contigs share a larger region of sequence identity, with the first 10,652 bases of NT_023464 being 99% identical to a region inside NT_008902. This difference could arise from an error in Cel assembly, caused by inability to distinguish two regions with very high sequence identity, or from insufficient coverage of the regions by high quality sequence reads. Alternatively, an HGP-nr error could arise from assembly limitations or insufficient high quality coverage of the single region endpoints, or from erroneous mapping of the single region to two chromosomes.

Searches for DNA-binding protein motifs

Compared with other organisms^{6,7}, the characterization of DNA-binding protein motifs in humans has been limited until recently by the paucity of sequence data for noncoding DNA. Although the draft human sequence has overcome this limitation, the large size of the human sequence and the existence of active regulatory sites located far from the genes they regulate⁸ pose new challenges to binding-site analysis. To assess how much they are enriched in the neighbourhood of genes in humans, we searched for motifs for two DNA-binding proteins (EGR-1 and CRX) in 4-kb upstream sequences of 3,352 genes and compared their abundances with those found in random 4-kb sequences and available positive control upstream sequences known or likely to contain binding sites (see Supplementary Information). We focused on HGP-nr sequence rather than Cel for upstream sequence because more annotation for genes was available, but used both HGP-all phase 3 and Cel for random sequences.

The human zinc finger DNA-binding protein EGR-1 (homologous to mouse Zif268) is induced by growth factors and nerve cell depolarization and is involved in cellular proliferation and differentiation⁹. We developed weight matrices that variously took into account a recognized nine-base consensus binding site (GCG[G/T]GGGCG¹⁰), *in vitro* selection data indicating a modified consensus and specific interactions with flanking bases¹¹, and our own double-stranded DNA array binding data, which indicates that the middle three bases of the site do not independently affect binding specificity¹². Figure 3a shows the results of a search for EGR-1 sites using matrix M1EGR-1, which looks for nine-base sites and uses a scoring threshold that allows only sites with small deviations from the consensus (see Supplementary Information). We found that upstream regions were significantly enriched for EGR-1 sites compared with random sequences, but not significantly different from 17 positive controls (Table 1). However, EGR-1 sites are GC-rich and therefore could occur more frequently by chance in

CpG islands upstream of genes. We therefore used a second matrix, M2EGR-1, which took into account the flanking bases of the nine-base site and compensated for the local GC content of potential sites when calculating scores (see Supplementary Information). The results of this search still indicate over-representation of EGR-1 sites in upstream regions (Fig. 3a, Table 1).

But not all DNA binding motifs show this pattern. We also searched for binding motifs for the photoreceptor homeobox transcription factor¹³ CRX. We generated a matrix, M1CRX, for CRX which indicated that CRX binding motifs are significantly under-represented in upstream regions compared with random sequence (Fig. 3b, Table 1). We are currently inspecting 67 CRX-regulated genes as potential positive controls. The only two inspected in detail so far, *RHO* (which encodes rhodopsin) and *PDC* (which encodes phosducin), are found to have three and two upstream sites, respectively. Although each of these is significantly higher than the number of sites found in the 3,352 upstream regions or random controls, the number of positive controls examined is too low to draw general conclusions (see Supplementary Information). These preliminary results suggest that DNA binding protein motifs have different statistical representation in upstream regions

compared with other sequence, but that over-representation in these regions does not necessarily indicate regulation by the protein. We are continuing our investigations with larger sets of upstream sequences (see Supplementary Information).

Although there are similar average numbers of DNA-binding sites in random sequences from Cel and HGP-all, they are significantly different (Table 1). This probably reflects subtle differences in the sequences that could be detected because such large samples were compared (83,800 for HGP-all and 33,520 for Cel). The differences between upstream sequence and random sequence were significant regardless of whether random sequence came from HGP-all or Cel.

Discussion

The HGP-nr and Cel draft genome assemblies are similar in size, contain comparable numbers of unique sequences (which we analysed as unique 15-nucleotide stretches, cu-15s), and exhibit similar statistics for sample candidate DNA protein-binding motifs. Some differences emerge at a detailed level. Sequence content aside, the assemblies are also packaged differently. Contigs in each exhibit different size and gap distributions (Figs 1 and 2), and HGP presents more stages of assembly by providing four phases of sequence data compared with the single Cel Human Fragments database. More annotation is also available at HGP. We expect all these differences to diminish as assemblies become more complete and comprehensive.

The complete human genome presents us with challenges on several levels. First, the analysis of 3 Gb of human sequence requires increased computer resources compared with analysis of smaller genomes, although this need can be addressed to an extent through modest computational goals and careful planning. For instance, a suffix array is a fast and convenient way of finding unique sub-sequences in a set of sequences¹⁴, but in our hands it requires 12 bytes of RAM per base pair of sequence when the number of sequences is large. The computer program described above required less than 300 Mbytes of RAM and could analyse cu-15s in the 4.8-Gb HGP-all in eight hours on a high-end conventional PC (see Supplementary Information). A second challenge is the fact that the genome sequences are constantly being updated, so programs that analyse them must be modified and rerun frequently to include the latest information and to accommodate new packaging and annotation. But this challenge brings with it the promise of rapid progress and new opportunity.

The broadest challenge will be the development of algorithms and platforms to meet the new computational needs that will arise as a

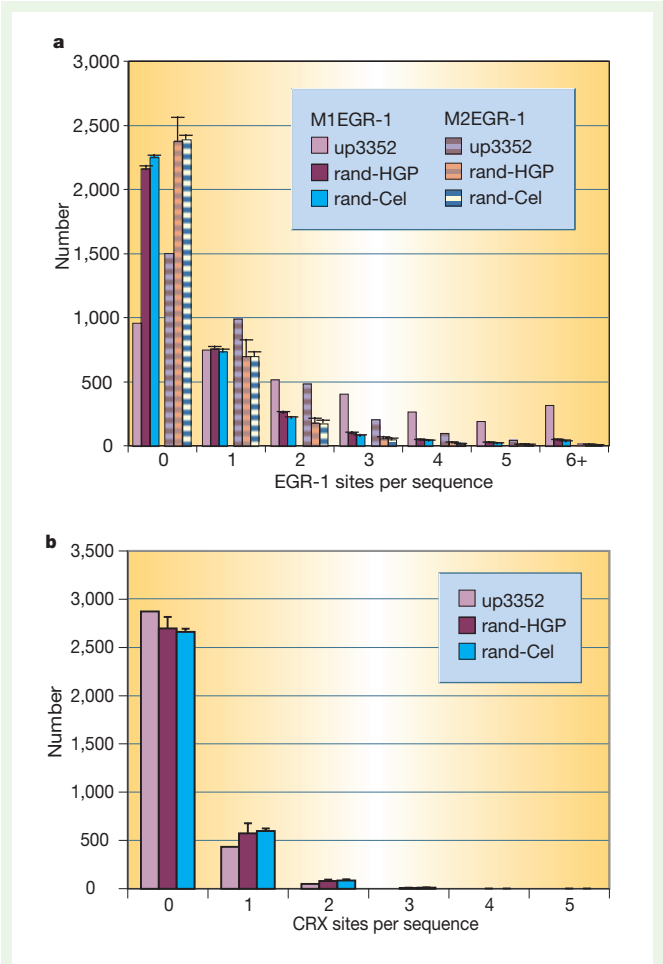


Figure 3 Numbers of DNA-binding sites per sequence from searches of upstream regions in sets of 3,352 4,000-bp sequences as determined by three different DNA-binding protein motif matrices. Sequence set abbreviations up3352, rand-HGP and rand-Cel are as in Table 1. up3352 bars represent counts of upstream sequences containing indicated numbers of sites. rand-HGP (rand-Cel) bars represent averages of counts of random sequences over 25 (10) selections of 3,352 sequences. Error bars for rand-HGP (rand-Cel) represent s.d. of the 25 (10) sets of counts. **a**, Sites per sequence for EGR-1 binding sites as determined by matrices M1EGR-1 and M2EGR-2. **b**, CRX sites as determined by matrix M1CRX.

Table 1 DNA-binding protein binding sites				
Sequence set	Sites per sequence	Sequence set comparisons		
		up3352	rand-HGP	rand-Cel
Matrix M1EGR-1				
up+c	2.35 ± 1.80	0.8140	1.4 × 10 ^{-7*}	2.2 × 10 ^{-10*}
up3352	2.21 ± 2.45		0*	0*
rand-HGP	0.64 ± 1.34			1.1 × 10 ^{-25*}
rand-Cel	0.55 ± 1.17			
Matrix M2EGR-1				
up+c	1.12 ± 0.99	0.6951	0.0032*	0.0002*
up3352	1.00 ± 1.26		5.5 × 10 ^{-240*}	0*
rand-HGP	0.42 ± 0.98			0.0004*
rand-Cel	0.40 ± 0.80			
Matrix M1CRX				
up3352	0.16 ± 0.40		1.4 × 10 ^{-14*}	7.0 × 10 ^{-19*}
rand-HGP	0.22 ± 0.48			7.8 × 10 ^{-6*}
rand-Cel	0.24 ± 0.50			

Average number ± s.d. of DNA-binding protein motifs found by different motif matrices in sets of 4,000-bp genomic sequences, and comparisons of average numbers found for different sets. Motif matrices M1EGR-1 and M2EGR-1 for EGR-1 binding sites, and M1CRX for CRX binding sites, are described in the text. Sequence sets: up+c, upstream regions of 17 EGR-1 positive control genes (see Supplementary Information); up3352, upstream regions extracted from HGP-nr sequence for 3,352 genes (see text); rand-HGP, 25 sets of 3,352 4,000-bp regions randomly extracted from HGP-all phase 3 sequence; rand-Cel, 10 sets of 3,352 4,000-bp regions randomly extracted from Cel sequence. Sequence set comparisons: t-test probabilities that average sites/sequence are equal for different sequence sets. Asterisk, statistically significant comparison (t-test *P* < 0.05).

result of the human genome sequence. For instance, any single 9–11-mer will occur 1,450–23,000 times in 6 Gb of DNA by chance alone, and therefore a 9–11-base recognition sequence for a protein such as EGR-1, which admits many variations, will occur many times more often. How do we distinguish actual regulatory sites? Focusing on regions near the 5' ends of genes may provide a first order approach, but the case of EGR-1 above suggests that this may not be very specific for functional sites. A promising approach is to look at regions of the upstream sequences that are conserved in other mammals^{15,16}. Similar genome-wide comparison requirements are arising in other contexts. For instance, new mutations have accumulated in the human population at a rate of 1–100 mutations per generation over the past 5,000 generations¹⁷. This means that among the 6 billion humans alive today, there is a reasonable chance that all possible single nucleotide polymorphisms exist for each of the approximately 3 billion base pairs of the human genome. It is becoming clear that costs must be reduced to study effectively the association of traits with polymorphisms¹⁸ or haplotypes¹⁹, but moving from population associations to accurate assessment of individual variations is likely to require sequence analysis of the two copies of the three billion base pairs in any of the six billion humans who desire it.

We remain optimistic that these challenges will be met. Through remarkable miniaturization and quality improvements, the cost of genomics has been dropping roughly twofold every 18 months for decades, paralleling the trend for computing²⁰. In a few years' time we may be able to read the 6 billion DNA base pairs in a human cell almost as easily and inexpensively as we can read a similar number of bits in a CD or DVD today. Today's milestone—two human genome sequence drafts—foreshadows a future in which resequencing and comparison of entire mammalian genomes will be routine operations for biology laboratories. □

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Maglott, D. R., Katz, K. S., Sicotte, H. & Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* **28**, 126–128 (2000).
4. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

5. Tatusova, T. A. & Madden, T. L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174**, 247–250 (1999).
6. McGuire, A. M., Hughes, J. D. & Church, G. M. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* **10**, 744–757 (2000).
7. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296**, 1205–1214 (2000).
8. Blackwood, E. M. & Kadonaga, J. T. Going the distance: a current view of enhancer action. *Science* **281**, 60–63 (1998).
9. Sukhatme, V. P. *et al.* A zinc finger-encoding gene coregulated with c-fos during growth and differentiation, and after cellular depolarization. *Cell* **53**, 37–43 (1988).
10. Christy, B. & Nathans, D. DNA binding site of the growth factor-inducible protein Zif268. *Proc. Natl Acad. Sci. USA* **86**, 8737–8741 (1989).
11. Swirnoff, A. H. & Milbrandt, J. DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Mol. Cell. Biol.* **15**, 2275–2287 (1995).
12. Bulky, M. L. Development and application of microarray technologies for the highly parallel analysis of the sequence specificity of DNA binding proteins. Dissertation (Harvard Medical School, Boston, 2000).
13. Livesey, F. J., Furukawa, T., Steffen, M. A., Church, G. M. & Cepko, C. L. Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene Crx. *Curr. Biol.* **10**, 301–310 (2000).
14. Gusfield, D. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology* (Cambridge Univ Press, Cambridge, 1997).
15. Lipman, D. J. Making (anti)sense of non-coding sequence conservation. *Nucleic Acids Res.* **25**, 3580–3583 (1997).
16. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.* **26**, 225–228 (2000).
17. Eyre-Walker, A. & Keightley, P. D. High genomic deleterious mutation rates in hominids. *Nature* **397**, 344–347 (1999).
18. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
19. Davidson, S. Research suggests importance of haplotypes over SNPs. *Nature Biotechnol.* **18**, 1134–1135 (2000).
20. Moore, G. M. Cramming more components onto integrated circuits. *Electron. Mag.* **38**, 114–117 (1965).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature. It is also available at <http://arep.med.harvard.edu>.

Acknowledgements

We thank J. Seidman, C. Cepko, S. Blackshaw and Y. Grad for information on genes considered in the paper; E. Lander, C. Venter, K. Pruitt, J. Ostell and J. Hughes for comments on this manuscript and information on genome sequences; P. Johnson and T. Chen for software assistance; and J. McPhee and M. Fanjoy for system support. We also thank the Lipper Foundation, DOE (G.M.C.), and the Paul & Daisy Soros Fellowship for New Americans (A.D.) for funding.

Correspondence should be addressed to G.M.C. (e-mail: church@arep.med.harvard.edu).

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century^{1–3} sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

The last quarter of a century has been marked by a relentless drive to decipher first genes and then entire genomes, spawning the field of genomics. The fruits of this work already include the genome sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, one fungus, two animals and one plant.

Here we report the results of a collaboration involving 20 groups from the United States, the United Kingdom, Japan, France, Germany and China to produce a draft sequence of the human genome. The draft genome sequence was generated from a physical map covering more than 96% of the euchromatic part of the human genome and, together with additional sequence in public databases, it covers about 94% of the human genome. The sequence was produced over a relatively short period, with coverage rising from about 10% to more than 90% over roughly fifteen months. The sequence data have been made available without restriction and updated daily throughout the project. The task ahead is to produce a finished sequence, by closing all gaps and resolving all ambiguities. Already about one billion bases are in final form and the task of bringing the vast majority of the sequence to this standard is now straightforward and should proceed rapidly.

The sequence of the human genome is of interest in several respects. It is the largest genome to be extensively sequenced so far, being 25 times as large as any previously sequenced genome and eight times as large as the sum of all such genomes. It is the first vertebrate genome to be extensively sequenced. And, uniquely, it is the genome of our own species.

Much work remains to be done to produce a complete finished sequence, but the vast trove of information that has become available through this collaborative effort allows a global perspective on the human genome. Although the details will change as the sequence is finished, many points are already clear.

- The genomic landscape shows marked variation in the distribution of a number of features, including genes, transposable elements, GC content, CpG islands and recombination rate. This gives us important clues about function. For example, the developmentally important HOX gene clusters are the most repeat-poor regions of the human genome, probably reflecting the very complex

coordinate regulation of the genes in the clusters.

- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.

- The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.

- Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from transposable elements.

- Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall activity of such elements in the hominid lineage. DNA transposons appear to have become completely inactive and long-terminal repeat (LTR) retrotransposons may also have done so.

- The pericentromeric and subtelomeric regions of chromosomes are filled with large recent segmental duplications of sequence from elsewhere in the genome. Segmental duplication is much more frequent in humans than in yeast, fly or worm.

- Analysis of the organization of Alu elements explains the long-standing mystery of their surprising genomic distribution, and suggests that there may be strong selection in favour of preferential retention of Alu elements in GC-rich regions and that these 'selfish' elements may benefit their human hosts.

- The mutation rate is about twice as high in male as in female meiosis, showing that most mutation occurs in males.

- Cytogenetic analysis of the sequenced clones confirms suggestions that large GC-poor regions are strongly correlated with 'dark G-bands' in karyotypes.

- Recombination rates tend to be much higher in distal regions (around 20 megabases (Mb)) of chromosomes and on shorter chromosome arms in general, in a pattern that promotes the occurrence of at least one crossover per chromosome arm in each meiosis.

- More than 1.4 million single nucleotide polymorphisms (SNPs) in the human genome have been identified. This collection should allow the initiation of genome-wide linkage disequilibrium mapping of the genes in the human population.

In this paper, we start by presenting background information on the project and describing the generation, assembly and evaluation of the draft genome sequence. We then focus on an initial analysis of the sequence itself: the broad chromosomal landscape; the repeat elements and the rich palaeontological record of evolutionary and biological processes that they provide; the human genes and proteins and their differences and similarities with those of other

Genome Sequencing Centres (Listed in order of total genomic sequence contributed, with a partial list of personnel. A full list of contributors at each centre is available as Supplementary Information.)

Whitehead Institute for Biomedical Research, Center for Genome Research: Eric S. Lander^{1*}, Lauren M. Linton¹, Bruce Birren^{1*}, Chad Nusbaum^{1*}, Michael C. Zody^{1*}, Jennifer Baldwin¹, Keri Devon¹, Ken Dewar¹, Michael Doyle¹, William FitzHugh^{1*}, Roel Funke¹, Diane Gage¹, Katrina Harris¹, Andrew Heaford¹, John Howland¹, Lisa Kann¹, Jessica Lehoczy¹, Rosie LeVine¹, Paul McEwan¹, Kevin McKernan¹, James Meldrim¹, Jill P. Mesirov^{1*}, Cher Miranda¹, William Morris¹, Jerome Naylor¹, Christina Raymond¹, Mark Rosetti¹, Ralph Santos¹, Andrew Sheridan¹, Carrie Sougnéz¹, Nicole Stange-Thomann¹, Nikola Stojanovic¹, Aravind Subramanian¹ & Dudley Wyman¹

The Sanger Centre: Jane Rogers², John Sulston^{2*}, Rachael Ainscough², Stephan Beck², David Bentley², John Burton², Christopher Clee², Nigel Carter², Alan Coulson², Rebecca Deadman², Panos Deloukas², Andrew Dunham², Ian Dunham², Richard Durbin^{2*}, Lisa French², Darren Grafham², Simon Gregory², Tim Hubbard^{2*}, Sean Humphray², Adrienne Hunt², Matthew Jones², Christine Lloyd², Amanda McMurray², Lucy Matthews², Simon Mercer², Sarah Milne², James C. Mullikin^{2*}, Andrew Mungall², Robert Plumb², Mark Ross², Ratna Shownkeen² & Sarah Sims²

Washington University Genome Sequencing Center: Robert H. Waterston^{3*}, Richard K. Wilson³, LaDeana W. Hillier^{3*}, John D. McPherson³, Marco A. Marra³, Elaine R. Mardis³, Lucinda A. Fulton³, Asif T. Chinwalla^{3*}, Kymberlie H. Pepin³, Warren R. Gish³, Stephanie L. Chissole³, Michael C. Wendl³, Kim D. Delehaunty³, Tracie L. Miner³, Andrew Delehaunty³, Jason B. Kramer³, Lisa L. Cook³, Robert S. Fulton³, Douglas L. Johnson³, Patrick J. Minx³ & Sandra W. Clifton³

US DOE Joint Genome Institute: Trevor Hawkins⁴, Elbert Branscomb⁴, Paul Predki⁴, Paul Richardson⁴, Sarah Wenning⁴, Tom Slezak⁴, Norman Doggett⁴, Jan-Fang Cheng⁴, Anne Olsen⁴, Susan Lucas⁴, Christopher Elkin⁴, Edward Uberbacher⁴ & Marvin Frazier⁴

Baylor College of Medicine Human Genome Sequencing Center: Richard A. Gibbs^{5*}, Donna M. Muzny⁵, Steven E. Scherer⁵, John B. Bouck^{5*}, Erica J. Sodergren⁵, Kim C. Worley^{5*}, Catherine M. Rives⁵, James H. Gorrell⁵, Michael L. Metzker⁵, Susan L. Naylor⁶, Raju S. Kucherlapati⁷, David L. Nelson, & George M. Weinstock⁸

RIKEN Genomic Sciences Center: Yoshiyuki Sakaki⁹, Asao Fujiyama⁹, Masahira Hattori⁹, Tetsushi Yada⁹, Atsushi Toyoda⁹, Takehiko Itoh⁹, Chiharu Kawagoe⁹, Hidemi Watanabe⁹, Yasushi Totoki⁹ & Todd Taylor⁹

Genoscope and CNRS UMR-8030: Jean Weissenbach¹⁰, Roland Heilig¹⁰, William Saurin¹⁰, Francois Artiguenave¹⁰, Philippe Brottier¹⁰, Thomas Bruls¹⁰, Eric Pelletier¹⁰, Catherine Robert¹⁰ & Patrick Wincker¹⁰

GTC Sequencing Center: Douglas R. Smith¹¹, Lynn Doucette-Stamm¹¹, Marc Rubenfield¹¹, Keith Weinstock¹¹, Hong Mei Lee¹¹ & JoAnn Dubois¹¹

Department of Genome Analysis, Institute of Molecular

Biotechnology: André Rosenthal¹², Matthias Platzer¹², Gerald Nyakatura¹², Stefan Taudien¹² & Andreas Rump¹²

Beijing Genomics Institute/Human Genome Center: Huanming Yang¹³, Jun Yu¹³, Jian Wang¹³, Guyang Huang¹⁴ & Jun Gu¹⁵

Multimegabase Sequencing Center, The Institute for Systems Biology: Leroy Hood¹⁶, Lee Rowen¹⁶, Anup Madan¹⁶ & Shizhen Qin¹⁶

Stanford Genome Technology Center: Ronald W. Davis¹⁷, Nancy A. Federspiel¹⁷, A. Pia Abola¹⁷ & Michael J. Proctor¹⁷

Stanford Human Genome Center: Richard M. Myers¹⁸, Jeremy Schmutz¹⁸, Mark Dickson¹⁸, Jane Grimwood¹⁸ & David R. Cox¹⁸

University of Washington Genome Center: Maynard V. Olson¹⁹, Rajinder Kaul¹⁹ & Christopher Raymond¹⁹

Department of Molecular Biology, Keio University School of Medicine: Nobuyoshi Shimizu²⁰, Kazuhiko Kawasaki²⁰ & Shinsei Minoshima²⁰

University of Texas Southwestern Medical Center at Dallas: Glen A. Evans²¹, Maria Athanasiou²¹ & Roger Schultz²¹

University of Oklahoma's Advanced Center for Genome Technology: Bruce A. Roe²², Feng Chen²² & Huaqin Pan²²

Max Planck Institute for Molecular Genetics: Juliane Ramser²³, Hans Lehrach²³ & Richard Reinhardt²³

Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center: W. Richard McCombie²⁴, Melissa de la Bastide²⁴ & Neilay Dedhia²⁴

GBF—German Research Centre for Biotechnology: Helmut Blöcker²⁵, Klaus Hornischer²⁵ & Gabriele Nordsiek²⁵

*** Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings):** Richa Agarwala²⁶, L. Aravind²⁶, Jeffrey A. Bailey²⁷, Alex Bateman², Serafim Batzoglou¹, Ewan Birney²⁸, Peer Bork^{29,30}, Daniel G. Brown¹, Christopher B. Burge³¹, Lorenzo Cerutti²⁸, Hsiu-Chuan Chen²⁶, Deanna Church²⁶, Michele Clamp², Richard R. Copley³⁰, Tobias Doerks^{29,30}, Sean R. Eddy³², Evan E. Eichler²⁷, Terrence S. Furey³³, James Galagan¹, James G. R. Gilbert², Cyrus Harmon³⁴, Yoshihide Hayashizaki³⁵, David Haussler³⁶, Henning Hermjakob²⁸, Karsten Hokamp³⁷, Wonhee Jang²⁶, L. Steven Johnson³², Thomas A. Jones³², Simon Kasif³⁸, Arek Kasprzyk²⁸, Scot Kennedy³⁹, W. James Kent⁴⁰, Paul Kitts²⁶, Eugene V. Koonin²⁶, Ian Korf³, David Kulp³⁴, Doron Lancet⁴¹, Todd M. Lowe⁴², Aoife McLysaght³⁷, Tarjei Mikkelsen³⁸, John V. Moran⁴³, Nicola Mulder²⁸, Victor J. Pollara¹, Chris P. Ponting⁴⁴, Greg Schuler²⁶, Jörg Schultz³⁰, Guy Slater²⁸, Arian F. A. Smit⁴⁵, Elia Stupka²⁸, Joseph Szustakowski³⁸, Danielle Thierry-Mieg²⁶, Jean Thierry-Mieg²⁶, Lukas Wagner²⁶, John Wallis³, Raymond Wheeler³⁴, Alan Williams³⁴, Yuri I. Wolf²⁶, Kenneth H. Wolfe³⁷, Shiaw-Pyng Yang³ & Ru-Fang Yeh³¹

Scientific management: National Human Genome Research Institute, US National Institutes of Health: Francis Collins^{46*}, Mark S. Guyer⁴⁶, Jane Peterson⁴⁶, Adam Felsenfeld^{46*} & Kris A. Wetterstrand⁴⁶, **Office of Science, US Department of Energy:** Aristides Patrinos⁴⁷, **The Wellcome Trust:** Michael J. Morgan⁴⁸

organisms; and the history of genomic segments. (Comparisons are drawn throughout with the genomes of the budding yeast *Saccharomyces cerevisiae*, the nematode worm *Caenorhabditis elegans*, the fruitfly *Drosophila melanogaster* and the mustard weed *Arabidopsis thaliana*; we refer to these for convenience simply as yeast, worm, fly and mustard weed.) Finally, we discuss applications of the sequence to biology and medicine and describe next steps in the project. A full description of the methods is provided as Supplementary Information on *Nature's* web site (<http://www.nature.com>).

We recognize that it is impossible to provide a comprehensive analysis of this vast dataset, and thus our goal is to illustrate the range of insights that can be gleaned from the human genome and thereby to sketch a research agenda for the future.

Background to the Human Genome Project

The Human Genome Project arose from two key insights that emerged in the early 1980s: that the ability to take global views of genomes could greatly accelerate biomedical research, by allowing researchers to attack problems in a comprehensive and unbiased fashion; and that the creation of such global views would require a communal effort in infrastructure building, unlike anything previously attempted in biomedical research. Several key projects helped to crystallize these insights, including:

- (1) The sequencing of the bacterial viruses Φ X174^{4,5} and lambda⁶, the animal virus SV40⁷ and the human mitochondrion⁸ between 1977 and 1982. These projects proved the feasibility of assembling small sequence fragments into complete genomes, and showed the value of complete catalogues of genes and other functional elements.
- (2) The programme to create a human genetic map to make it possible to locate disease genes of unknown function based solely on their inheritance patterns, launched by Botstein and colleagues in 1980 (ref. 9).
- (3) The programmes to create physical maps of clones covering the yeast¹⁰ and worm¹¹ genomes to allow isolation of genes and regions based solely on their chromosomal position, launched by Olson and Sulston in the mid-1980s.

- (4) The development of random shotgun sequencing of complementary DNA fragments for high-throughput gene discovery by Schimmel¹² and Schimmel and Sutcliffe¹³, later dubbed expressed sequence tags (ESTs) and pursued with automated sequencing by Venter and others^{14–20}.

The idea of sequencing the entire human genome was first proposed in discussions at scientific meetings organized by the US Department of Energy and others from 1984 to 1986 (refs 21, 22). A committee appointed by the US National Research Council endorsed the concept in its 1988 report²³, but recommended a broader programme, to include: the creation of genetic, physical and sequence maps of the human genome; parallel efforts in key model organisms such as bacteria, yeast, worms, flies and mice; the development of technology in support of these objectives; and research into the ethical, legal and social issues raised by human genome research. The programme was launched in the US as a joint effort of the Department of Energy and the National Institutes of Health. In other countries, the UK Medical Research Council and the Wellcome Trust supported genomic research in Britain; the Centre d'Etude du Polymorphisme Humain and the French Muscular Dystrophy Association launched mapping efforts in France; government agencies, including the Science and Technology Agency and the Ministry of Education, Science, Sports and Culture supported genomic research efforts in Japan; and the European Community helped to launch several international efforts, notably the programme to sequence the yeast genome. By late 1990, the Human Genome Project had been launched, with the creation of genome centres in these countries. Additional participants subsequently joined the effort, notably in Germany and China. In addition, the Human Genome Organization (HUGO) was founded to provide a forum for international coordination of genomic research. Several books^{24–26} provide a more comprehensive discussion of the genesis of the Human Genome Project.

Through 1995, work progressed rapidly on two fronts (Fig. 1). The first was construction of genetic and physical maps of the human and mouse genomes^{27–31}, providing key tools for identification of disease genes and anchoring points for genomic sequence. The second was sequencing of the yeast³² and worm³³ genomes, as

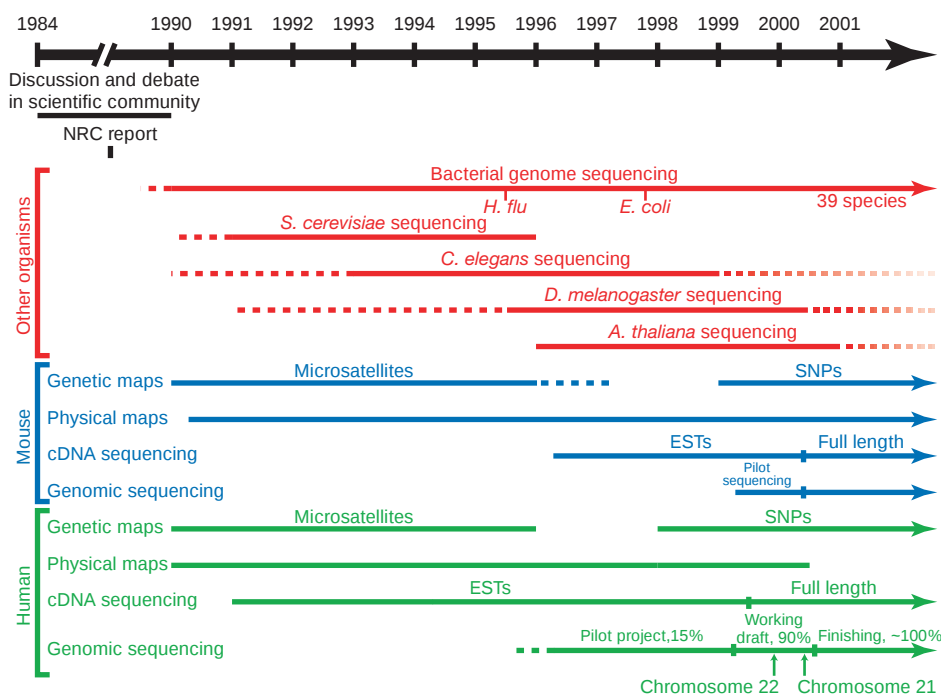


Figure 1 Timeline of large-scale genomic analyses. Shown are selected components of work on several non-vertebrate model organisms (red), the mouse (blue) and the human

(green) from 1990; earlier projects are described in the text. SNPs, single nucleotide polymorphisms; ESTs, expressed sequence tags.

well as targeted regions of mammalian genomes^{34–37}. These projects showed that large-scale sequencing was feasible and developed the two-phase paradigm for genome sequencing. In the first, ‘shotgun’ phase, the genome is divided into appropriately sized segments and each segment is covered to a high degree of redundancy (typically, eight- to tenfold) through the sequencing of randomly selected subfragments. The second is a ‘finishing’ phase, in which sequence gaps are closed and remaining ambiguities are resolved through directed analysis. The results also showed that complete genomic sequence provided information about genes, regulatory regions and chromosome structure that was not readily obtainable from cDNA studies alone.

In 1995, genome scientists considered a proposal³⁸ that would have involved producing a draft genome sequence of the human genome in a first phase and then returning to finish the sequence in a second phase. After vigorous debate, it was decided that such a plan was premature for several reasons. These included the need first to prove that high-quality, long-range finished sequence could be produced from most parts of the complex, repeat-rich human genome; the sense that many aspects of the sequencing process were still rapidly evolving; and the desirability of further decreasing costs.

Instead, pilot projects were launched to demonstrate the feasibility of cost-effective, large-scale sequencing, with a target completion date of March 1999. The projects successfully produced finished sequence with 99.99% accuracy and no gaps³⁹. They also introduced bacterial artificial chromosomes (BACs)⁴⁰, a new large-insert cloning system that proved to be more stable than the cosmids and yeast artificial chromosomes (YACs)⁴¹ that had been used previously. The pilot projects drove the maturation and convergence of sequencing strategies, while producing 15% of the human genome sequence. With successful completion of this phase, the human genome sequencing effort moved into full-scale production in March 1999.

The idea of first producing a draft genome sequence was revived at this time, both because the ability to finish such a sequence was no longer in doubt and because there was great hunger in the scientific community for human sequence data. In addition, some scientists favoured prioritizing the production of a draft genome sequence over regional finished sequence because of concerns about commercial plans to generate proprietary databases of human sequence that might be subject to undesirable restrictions on use^{42–44}.

The consortium focused on an initial goal of producing, in a first production phase lasting until June 2000, a draft genome sequence covering most of the genome. Such a draft genome sequence, although not completely finished, would rapidly allow investigators to begin to extract most of the information in the human sequence. Experiments showed that sequencing clones covering about 90% of the human genome to a redundancy of about four- to fivefold (‘half-shotgun’ coverage; see Box 1) would accomplish this^{45,46}. The draft genome sequence goal has been achieved, as described below.

The second sequence production phase is now under way. Its aims are to achieve full-shotgun coverage of the existing clones during 2001, to obtain clones to fill the remaining gaps in the physical map, and to produce a finished sequence (apart from regions that cannot be cloned or sequenced with currently available techniques) no later than 2003.

Strategic issues

Hierarchical shotgun sequencing

Soon after the invention of DNA sequencing methods^{47,48}, the shotgun sequencing strategy was introduced^{49–51}; it has remained the fundamental method for large-scale genome sequencing^{52–54} for the past 20 years. The approach has been refined and extended to make it more efficient. For example, improved protocols for fragmenting and cloning DNA allowed construction of shotgun

libraries with more uniform representation. The practice of sequencing from both ends of double-stranded clones (‘double-barrelled’ shotgun sequencing) was introduced by Ansorge and others³⁷ in 1990, allowing the use of ‘linking information’ between sequence fragments.

The application of shotgun sequencing was also extended by applying it to larger and larger DNA molecules—from plasmids (~4 kilobases (kb)) to cosmid clones³⁷ (40 kb), to artificial chromosomes cloned in bacteria and yeast⁵⁵ (100–500 kb) and bacterial genomes⁵⁶ (1–2 megabases (Mb)). In principle, a genome of arbitrary size may be directly sequenced by the shotgun method, provided that it contains no repeated sequence and can be uniformly sampled at random. The genome can then be assembled using the simple computer science technique of ‘hashing’ (in which one detects overlaps by consulting an alphabetized look-up table of all *k*-letter words in the data). Mathematical analysis of the expected number of gaps as a function of coverage is similarly straightforward⁵⁷.

Practical difficulties arise because of repeated sequences and cloning bias. Small amounts of repeated sequence pose little problem for shotgun sequencing. For example, one can readily assemble typical bacterial genomes (about 1.5% repeat) or the euchromatic portion of the fly genome (about 3% repeat). By contrast, the human genome is filled (>50%) with repeated sequences, including interspersed repeats derived from transposable elements, and long genomic regions that have been duplicated in tandem, palindromic or dispersed fashion (see below). These include large duplicated segments (50–500 kb) with high sequence identity (98–99.9%), at which mispairing during recombination creates deletions responsible for genetic syndromes. Such features complicate the assembly of a correct and finished genome sequence.

There are two approaches for sequencing large repeat-rich genomes. The first is a whole-genome shotgun sequencing approach, as has been used for the repeat-poor genomes of viruses, bacteria and flies, using linking information and computational

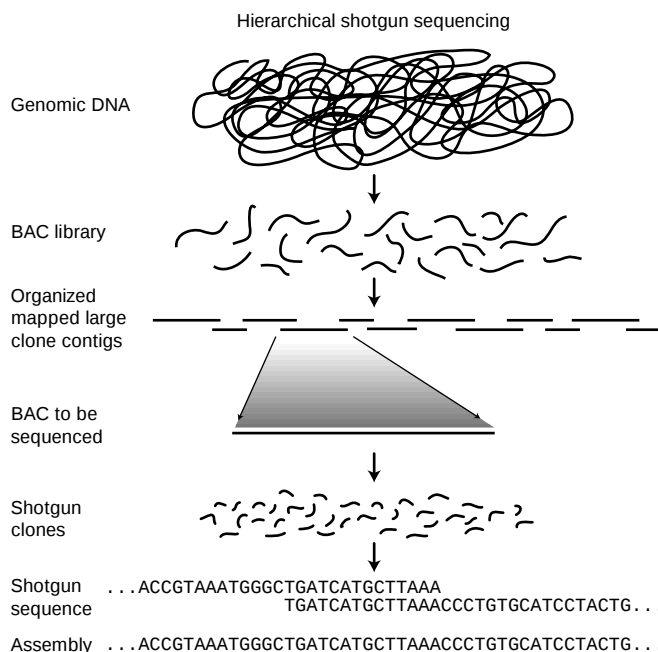


Figure 2 Idealized representation of the hierarchical shotgun sequencing strategy. A library is constructed by fragmenting the target genome and cloning it into a large-fragment cloning vector; here, BAC vectors are shown. The genomic DNA fragments represented in the library are then organized into a physical map and individual BAC clones are selected and sequenced by the random shotgun strategy. Finally, the clone sequences are assembled to reconstruct the sequence of the genome.

analysis to attempt to avoid misassemblies. The second is the 'hierarchical shotgun sequencing' approach (Fig. 2), also referred to as 'map-based', 'BAC-based' or 'clone-by-clone'. This approach involves generating and organizing a set of large-insert clones (typically 100–200 kb each) covering the genome and separately performing shotgun sequencing on appropriately chosen clones. Because the sequence information is local, the issue of long-range misassembly is eliminated and the risk of short-range misassembly is reduced. One caveat is that some large-insert clones may suffer rearrangement, although this risk can be reduced by appropriate quality-control measures involving clone fingerprints (see below).

The two methods are likely to entail similar costs for producing finished sequence of a mammalian genome. The hierarchical approach has a higher initial cost than the whole-genome approach, owing to the need to create a map of clones (about 1% of the total cost of sequencing) and to sequence overlaps between clones. On the other hand, the whole-genome approach is likely to require much greater work and expense in the final stage of producing a finished sequence, because of the challenge of resolving misassemblies. Both methods must also deal with cloning biases, resulting in under-representation of some regions in either large-insert or small-insert clone libraries.

There was lively scientific debate over whether the human genome sequencing effort should employ whole-genome or hierarchical shotgun sequencing. Weber and Myers⁵⁸ stimulated these discussions with a specific proposal for a whole-genome shotgun approach, together with an analysis suggesting that the method could work and be more efficient. Green⁵⁹ challenged these conclusions and argued that the potential benefits did not outweigh the likely risks.

In the end, we concluded that the human genome sequencing effort should employ the hierarchical approach for several reasons. First, it was prudent to use the approach for the first project to sequence a repeat-rich genome. With the hierarchical approach, the ultimate frequency of misassembly in the finished product would probably be lower than with the whole-genome approach, in which it would be more difficult to identify regions in which the assembly was incorrect.

Second, it was prudent to use the approach in dealing with an outbred organism, such as the human. In the whole-genome shotgun method, sequence would necessarily come from two different copies of the human genome. Accurate sequence assembly could be complicated by sequence variation between these two copies—both SNPs (which occur at a rate of 1 per 1,300 bases) and larger-scale structural heterozygosity (which has been documented in human chromosomes). In the hierarchical shotgun method, each large-insert clone is derived from a single haplotype.

Third, the hierarchical method would be better able to deal with inevitable cloning biases, because it would more readily allow targeting of additional sequencing to under-represented regions. And fourth, it was better suited to a project shared among members of a diverse international consortium, because it allowed work and responsibility to be easily distributed. As the ultimate goal has always been to create a high-quality, finished sequence to serve as a foundation for biomedical research, we reasoned that the advantages of this more conservative approach outweighed the additional cost, if any.

A biotechnology company, Celera Genomics, has chosen to incorporate the whole-genome shotgun approach into its own efforts to sequence the human genome. Their plan^{60,61} uses a mixed strategy, involving combining some coverage with whole-genome shotgun data generated by the company together with the publicly available hierarchical shotgun data generated by the International Human Genome Sequencing Consortium. If the raw sequence reads from the whole-genome shotgun component are made available, it may be possible to evaluate the extent to which the sequence of the human genome can be assembled without the need

for clone-based information. Such analysis may help to refine sequencing strategies for other large genomes.

Technology for large-scale sequencing

Sequencing the human genome depended on many technological improvements in the production and analysis of sequence data. Key innovations were developed both within and outside the Human Genome Project. Laboratory innovations included four-colour fluorescence-based sequence detection⁶², improved fluorescent dyes^{63–66}, dye-labelled terminators⁶⁷, polymerases specifically designed for sequencing^{68–70}, cycle sequencing⁷¹ and capillary gel electrophoresis^{72–74}. These studies contributed to substantial improvements in the automation, quality and throughput of collecting raw DNA sequence^{75,76}. There were also important advances in the development of software packages for the analysis of sequence data. The PHRED software package^{77,78} introduced the concept of assigning a 'base-quality score' to each base, on the basis of the probability of an erroneous call. These quality scores make it possible to monitor raw data quality and also assist in determining whether two similar sequences truly overlap. The PHRAP computer package (<http://bozeman.mbt.washington.edu/phrap.docs/phrap.html>) then systematically assembles the sequence data using the base-quality scores. The program assigns 'assembly-quality scores' to each base in the assembled sequence, providing an objective criterion to guide sequence finishing. The quality scores were based on and validated by extensive experimental data.

Another key innovation for scaling up sequencing was the development by several centres of automated methods for sample preparation. This typically involved creating new biochemical protocols suitable for automation, followed by construction of appropriate robotic systems.

Coordination and public data sharing

The Human Genome Project adopted two important principles with regard to human sequencing. The first was that the collaboration would be open to centres from any nation. Although potentially less efficient, in a narrow economic sense, than a centralized approach involving a few large factories, the inclusive approach was strongly favoured because we felt that the human genome sequence is the common heritage of all humanity and the work should transcend national boundaries, and we believed that scientific progress was best assured by a diversity of approaches. The collaboration was coordinated through periodic international meetings (referred to as 'Bermuda meetings' after the venue of the first three gatherings) and regular telephone conferences. Work was shared flexibly among the centres, with some groups focusing on particular chromosomes and others contributing in a genome-wide fashion.

The second principle was rapid and unrestricted data release. The centres adopted a policy that all genomic sequence data should be made publicly available without restriction within 24 hours of assembly^{79,80}. Pre-publication data releases had been pioneered in mapping projects in the worm¹¹ and mouse genomes^{30,81} and were prominently adopted in the sequencing of the worm, providing a direct model for the human sequencing efforts. We believed that scientific progress would be most rapidly advanced by immediate and free availability of the human genome sequence. The explosion of scientific work based on the publicly available sequence data in both academia and industry has confirmed this judgement.

Generating the draft genome sequence

Generating a draft sequence of the human genome involved three steps: selecting the BAC clones to be sequenced, sequencing them and assembling the individual sequenced clones into an overall draft genome sequence. A glossary of terms related to genome sequencing and assembly is provided in Box 1.

The draft genome sequence is a dynamic product, which is regularly updated as additional data accumulate en route to the

ultimate goal of a completely finished sequence. The results below are based on the map and sequence data available on 7 October 2000, except as otherwise noted. At the end of this section, we provide a brief update of key data.

Clone selection

The hierarchical shotgun method involves the sequencing of overlapping large-insert clones spanning the genome. For the Human Genome Project, clones were largely chosen from eight large-insert libraries containing BAC or P1-derived artificial chromosome (PAC) clones (Table 1; refs 82–88). The libraries were made by

partial digestion of genomic DNA with restriction enzymes. Together, they represent around 65-fold coverage (redundant sampling) of the genome. Libraries based on other vectors, such as cosmids, were also used in early stages of the project.

The libraries (Table 1) were prepared from DNA obtained from anonymous human donors in accordance with US Federal Regulations for the Protection of Human Subjects in Research (45CFR46) and following full review by an Institutional Review Board. Briefly, the opportunity to donate DNA for this purpose was broadly advertised near the two laboratories engaged in library

Box 1

Genome glossary

Sequence

Raw sequence Individual unassembled sequence reads, produced by sequencing of clones containing DNA inserts.

Paired-end sequence Raw sequence obtained from both ends of a cloned insert in any vector, such as a plasmid or bacterial artificial chromosome.

Finished sequence Complete sequence of a clone or genome, with an accuracy of at least 99.99% and no gaps.

Coverage (or depth) The average number of times a nucleotide is represented by a high-quality base in a collection of random raw sequence. Operationally, a 'high-quality base' is defined as one with an accuracy of at least 99% (corresponding to a PHRED score of at least 20).

Full shotgun coverage The coverage in random raw sequence needed from a large-insert clone to ensure that it is ready for finishing; this varies among centres but is typically 8–10-fold. Clones with full shotgun coverage can usually be assembled with only a handful of gaps per 100 kb.

Half shotgun coverage Half the amount of full shotgun coverage (typically, 4–5-fold random coverage).

Clones

BAC clone Bacterial artificial chromosome vector carrying a genomic DNA insert, typically 100–200 kb. Most of the large-insert clones sequenced in the project were BAC clones.

Finished clone A large-insert clone that is entirely represented by finished sequence.

Full shotgun clone A large-insert clone for which full shotgun sequence has been produced.

Draft clone A large-insert clone for which roughly half-shotgun sequence has been produced. Operationally, the collection of draft clones produced by each centre was required to have an average coverage of fourfold for the entire set and a minimum coverage of threefold for each clone.

Predraft clone A large-insert clone for which some shotgun sequence is available, but which does not meet the standards for inclusion in the collection of draft clones.

Contigs and scaffolds

Contig The result of joining an overlapping collection of sequences or clones.

Scaffold The result of connecting contigs by linking information from paired-end reads from plasmids, paired-end reads from BACs, known messenger RNAs or other sources. The contigs in a scaffold are ordered and oriented with respect to one another.

Fingerprint clone contigs Contigs produced by joining clones inferred to overlap on the basis of their restriction digest fingerprints.

Sequenced-clone layout Assignment of sequenced clones to the physical map of fingerprint clone contigs.

Initial sequence contigs Contigs produced by merging overlapping sequence reads obtained from a single clone, in a process called sequence assembly.

Merged sequence contigs Contigs produced by taking the initial sequence contigs contained in overlapping clones and merging those found to overlap. These are also referred to simply as 'sequence contigs' where no confusion will result.

Sequence-contig scaffolds Scaffolds produced by connecting sequence contigs on the basis of linking information.

Sequenced-clone contigs Contigs produced by merging overlapping sequenced clones.

Sequenced-clone-contig scaffolds Scaffolds produced by joining sequenced-clone contigs on the basis of linking information.

Draft genome sequence The sequence produced by combining the information from the individual sequenced clones (by creating merged sequence contigs and then employing linking information to create scaffolds) and positioning the sequence along the physical map of the chromosomes.

N50 length A measure of the contig length (or scaffold length) containing a 'typical' nucleotide. Specifically, it is the maximum length L such that 50% of all nucleotides lie in contigs (or scaffolds) of size at least L .

Computer programs and databases

PHRED A widely used computer program that analyses raw sequence to produce a 'base call' with an associated 'quality score' for each position in the sequence. A PHRED quality score of X corresponds to an error probability of approximately $10^{-X/10}$. Thus, a PHRED quality score of 30 corresponds to 99.9% accuracy for the base call in the raw read.

PHRAP A widely used computer program that assembles raw sequence into sequence contigs and assigns to each position in the sequence an associated 'quality score', on the basis of the PHRED scores of the raw sequence reads. A PHRAP quality score of X corresponds to an error probability of approximately $10^{-X/10}$. Thus, a PHRAP quality score of 30 corresponds to 99.9% accuracy for a base in the assembled sequence.

GigAssembler A computer program developed during this project for merging the information from individual sequenced clones into a draft genome sequence.

Public sequence databases The three coordinated international sequence databases: GenBank, the EMBL data library and DDBJ.

Map features

STS Sequence tagged site, corresponding to a short (typically less than 500 bp) unique genomic locus for which a polymerase chain reaction assay has been developed.

EST Expressed sequence tag, obtained by performing a single raw sequence read from a random complementary DNA clone.

SSR Simple sequence repeat, a sequence consisting largely of a tandem repeat of a specific k -mer (such as $(CA)_{15}$). Many SSRs are polymorphic and have been widely used in genetic mapping.

SNP Single nucleotide polymorphism, or a single nucleotide position in the genome sequence for which two or more alternative alleles are present at appreciable frequency (traditionally, at least 1%) in the human population.

Genetic map A genome map in which polymorphic loci are positioned relative to one another on the basis of the frequency with which they recombine during meiosis. The unit of distance is centimorgans (cM), denoting a 1% chance of recombination.

Radiation hybrid (RH) map A genome map in which STSs are positioned relative to one another on the basis of the frequency with which they are separated by radiation-induced breaks. The frequency is assayed by analysing a panel of human–hamster hybrid cell lines, each produced by lethally irradiating human cells and fusing them with recipient hamster cells such that each carries a collection of human chromosomal fragments. The unit of distance is centirays (cR), denoting a 1% chance of a break occurring between two loci.

construction. Volunteers of diverse backgrounds were accepted on a first-come, first-taken basis. Samples were obtained after discussion with a genetic counsellor and written informed consent. The samples were made anonymous as follows: the sampling laboratory stripped all identifiers from the samples, applied random numeric labels, and transferred them to the processing laboratory, which then removed all labels and relabelled the samples. All records of the labelling were destroyed. The processing laboratory chose samples at random from which to prepare DNA and immortalized cell lines. Around 5–10 samples were collected for every one that was eventually used. Because no link was retained between donor and DNA sample, the identity of the donors for the libraries is not known, even by the donors themselves. A more complete description can be found at http://www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/human_subjects.html.

During the pilot phase, centres showed that sequence-tagged sites (STSs) from previously constructed genetic and physical maps could be used to recover BACs from specific regions. As sequencing expanded, some centres continued this approach, augmented with additional probes from flow sorting of chromosomes to obtain long-range coverage of specific chromosomes or chromosomal regions^{89–94}.

For the large-scale sequence production phase, a genome-wide physical map of overlapping clones was also constructed by systematic analysis of BAC clones representing 20-fold coverage of the human genome⁸⁶. Most clones came from the first three sections of the RPCI-11 library, supplemented with clones from sections of the

RPCI-13 and CalTech D libraries (Table 1). DNA from each BAC clone was digested with the restriction enzyme *HindIII*, and the sizes of the resulting fragments were measured by agarose gel electrophoresis. The pattern of restriction fragments provides a 'fingerprint' for each BAC, which allows different BACs to be distinguished and the degree of overlaps to be assessed. We used these restriction-fragment fingerprints to determine clone overlaps, and thereby assembled the BACs into fingerprint clone contigs.

The fingerprint clone contigs were positioned along the chromosomes by anchoring them with STS markers from existing genetic and physical maps. Fingerprint clone contigs were tied to specific STSs initially by probe hybridization and later by direct search of the sequenced clones. To localize fingerprint clone contigs that did not contain known markers, new STSs were generated and placed onto chromosomes⁹⁵. Representative clones were also positioned by fluorescence *in situ* hybridization (FISH) (ref. 86 and C. McPherson, unpublished).

We selected clones from the fingerprint clone contigs for sequencing according to various criteria. Fingerprint data were reviewed^{86,90} to evaluate overlaps and to assess clone fidelity (to bias against rearranged clones^{83,96}). STS content information and BAC end sequence information were also used^{91,92}. Where possible, we tried to select a minimally overlapping set spanning a region. However, because the genome-wide physical map was constructed concurrently with the sequencing, continuity in many regions was low in early stages. These small fingerprint clone contigs were nonetheless useful in identifying validated, nonredundant clones

Table 1 Key large-insert genome-wide libraries

Library name*	GenBank abbreviation	Vector type	Source DNA	Library segment or plate numbers	Enzyme digest	Average insert size (kb)	Total number of clones in library	Number of fingerprinted clones†	BAC-end sequence (ends/clones/clones with both ends sequenced)‡	Number of clones in genome layout§	Sequenced clones used in construction of the draft genome sequence		
											Number	Total bases (Mb)	Fraction of total from library
Caltech B	CTB	BAC	987SK cells	All	<i>HindIII</i>	120	74,496	16	2/1/1	528	518	66.7	0.016
Caltech C	CTC	BAC	Human sperm	All	<i>HindIII</i>	125	263,040	144	21,956/14,445/7,255	621	606	88.4	0.021
Caltech D1 (CITB-H1)	CTD	BAC	Human sperm	All	<i>HindIII</i>	129	162,432	49,833	403,589/226,068/156,631	1,381	1,367	185.6	0.043
Caltech D2 (CITB-E1)		BAC	Human sperm	All									
				2,501–2,565	<i>EcoRI</i>	202	24,960						
				2,566–2,671	<i>EcoRI</i>	182	46,326						
				3,000–3,253	<i>EcoRI</i>	142	97,536						
RPCI-1	RP1	PAC	Male, blood	All	<i>Mbol</i>	110	115,200	3,388		1,070	1,053	117.7	0.028
RPCI-3	RP3	PAC	Male, blood	All	<i>Mbol</i>	115	75,513			644	638	68.5	0.016
RPCI-4	RP4	PAC	Male, blood	All	<i>Mbol</i>	116	105,251			889	881	95.5	0.022
RPCI-5	RP5	PAC	Male, blood	All	<i>Mbol</i>	115	142,773			1,042	1,033	116.5	0.027
RPCI-11	RP11	BAC	Male, blood	All		178	543,797	267,931	379,773/243,764/134,110	19,405	19,145	3,165.0	0.743
				1	<i>EcoRI</i>	164	108,499						
				2	<i>EcoRI</i>	168	109,496						
				3	<i>EcoRI</i>	181	109,657						
				4	<i>EcoRI</i>	183	109,382						
				5	<i>Mbol</i>	196	106,763						
Total of top eight libraries							1,482,502	321,312	805,320/484,278/297,997	25,580	25,241	3,903.9	0.916
Total all libraries								354,510	812,594/488,017/100,775	30,445	29,298	4,260.5	1

* For the CalTech libraries⁹², see http://www.tree.caltech.edu/lib_status.html; for RPCI libraries⁸³, see <http://www.chori.org/bacpac/home.htm>.

† For the FPC map and fingerprinting^{84–86}, see http://genome.wustl.edu/gsc/human/human_database.shtml.

‡ The number of raw BAC end sequences (clones/ends/clones with both ends sequenced) available for use in human genome sequencing. Typically, for clones in which sequence was obtained from both ends, more than 95% of both end sequences contained at least 100 bp of nonrepetitive sequence. BAC-end sequencing of RPCI-11 and of the CalTech libraries was done at The Institute for Genomic Research, the California Institute of Technology and the University of Washington High Throughput Sequencing Center. The sources for the Table were <http://www.ncbi.nlm.nih.gov/genome/clone/BESstat.shtml> and refs 87, 88.

§ These are the clones in the sequenced-clone layout map (<http://genome.wustl.edu/gsc/human/Mapping/index.shtml>) that were pre-draft, draft or finished.

|| The number of sequenced clones used in the assembly. This number is less than that in the previous column owing to removal of a small number of obviously contaminated, combined or duplicated projects; in addition, not all of the clones from completed chromosomes 21 and 22 were included here because only the available finished sequence from those chromosomes was used in the assembly.

¶ The number reported is the total sequence from the clones indicated in the previous column. Potential overlap between clones was not removed here, but Ns were excluded.

that were used to 'seed' the sequencing of new regions. The small fingerprint clone contigs were extended or merged with others as the map matured.

The clones that make up the draft genome sequence therefore do not constitute a minimally overlapping set—there is overlap and redundancy in places. The cost of using suboptimal overlaps was justified by the benefit of earlier availability of the draft genome sequence data. Minimizing the overlap between adjacent clones would have required completing the physical map before undertaking large-scale sequencing. In addition, the overlaps between BAC clones provide a rich collection of SNPs. More than 1.4 million SNPs have already been identified from clone overlaps and other sequence comparisons⁹⁷.

Because the sequencing project was shared among twenty centres in six countries, it was important to coordinate selection of clones across the centres. Most centres focused on particular chromosomes or, in some cases, larger regions of the genome. We also maintained a clone registry to track selected clones and their progress. In later phases, the global map provided an integrated view of the data from all centres, facilitating the distribution of effort to maximize coverage of the genome. Before performing extensive sequencing on a

clone, several centres routinely examined an initial sample of 96 raw sequence reads from each subclone library to evaluate possible overlap with previously sequenced clones.

Sequencing

The selected clones were subjected to shotgun sequencing. Although the basic approach of shotgun sequencing is well established, the details of implementation varied among the centres. For example, there were differences in the average insert size of the shotgun libraries, in the use of single-stranded or double-stranded cloning vectors, and in sequencing from one end or both ends of each insert. Centres differed in the fluorescent labels employed and in the degree to which they used dye-primers or dye-terminators. The sequence detectors included both slab gel- and capillary-based devices. Detailed protocols are available on the web sites of many of the individual centres (URLs can be found at www.nhgri.nih.gov/genome_hub). The extent of automation also varied greatly among the centres, with the most aggressive automation efforts resulting in factory-style systems able to process more than 100,000 sequencing reactions in 12 hours (Fig. 3). In addition, centres differed in the amount of raw sequence data typically obtained for each clone (so-called half-shotgun, full shotgun and finished sequence). Sequence information from the different centres could be directly integrated despite this diversity, because the data were analysed by a common computational procedure. Raw sequence traces were processed and assembled with the PHRED and PHRAP software packages^{77,78} (P. Green, unpublished). All assembled contigs of more than 2 kb were deposited in public databases within 24 hours of assembly.

The overall sequencing output rose sharply during production (Fig. 4). Following installation of new sequence detectors beginning in June 1999, sequencing capacity and output rose approximately eightfold in eight months to nearly 7 million samples processed per month, with little or no drop in success rate (ratio of useable reads to attempted reads). By June 2000, the centres were producing raw sequence at a rate equivalent to onefold coverage of the entire human genome in less than six weeks. This corresponded to a continuous throughput exceeding 1,000 nucleotides per second, 24 hours per day, seven days per week. This scale-up resulted in a concomitant increase in the sequence available in the public databases (Fig. 4).

A version of the draft genome sequence was prepared on the basis of the map and sequence data available on 7 October 2000. For this version, the mapping effort had assembled the fingerprinted BACs into 1,246 fingerprint clone contigs. The sequencing effort had sequenced and assembled 29,298 overlapping BACs and other large-insert clones (Table 2), comprising a total length of 4.26 gigabases (Gb). This resulted from around 23 Gb of underlying raw shotgun sequence data, or about 7.5-fold coverage averaged across the genome (including both draft and finished sequence). The various contributions to the total amount of sequence deposited in the HTGS division of GenBank are given in Table 3.



Figure 3 The automated production line for sample preparation at the Whitehead Institute, Center for Genome Research. The system consists of custom-designed factory-style conveyor belt robots that perform all functions from purifying DNA from bacterial cultures through setting up and purifying sequencing reactions.

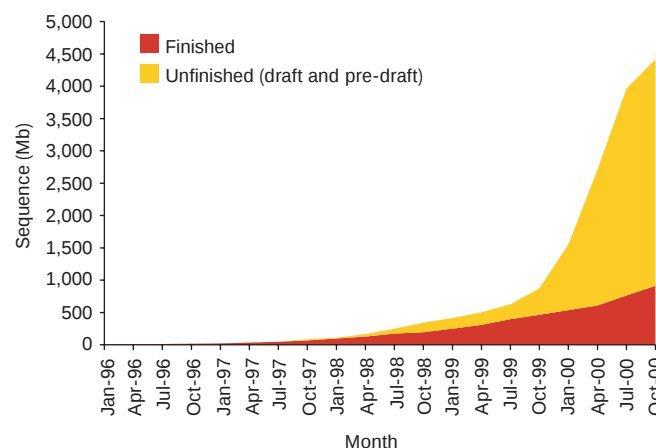


Figure 4 Total amount of human sequence in the High Throughput Genome Sequence (HTGS) division of GenBank. The total is the sum of finished sequence (red) and unfinished (draft plus pre-draft) sequence (yellow).

Table 2 Total genome sequence from the collection of sequenced clones, by sequence status

Sequence status	Number of clones	Total clone length (Mb)	Average number of sequence reads per kb*	Average sequence depth†	Total amount of raw sequence (Mb)
Finished	8,277	897	20–25	8–12	9,085
Draft	18,969	3,097	12	4.5	13,395
Predraft	2,052	267	6	2.5	667
Total					23,147

* The average number of reads per kb was estimated based on information provided by each sequencing centre. This number differed among sequencing centres, based on the actual protocols used.

† The average depth in high quality bases ($\geq 99\%$ accuracy) was estimated from information provided by each sequencing centre. The average varies among the centres, and the number may vary considerably for clones with the same sequencing status. For draft clones in the public databases (keyword: HTGS_draft), the number can be computed from the quality scores listed in the database entry.

By agreement among the centres, the collection of draft clones produced by each centre was required to have fourfold average sequence coverage, with no clone below threefold. (For this purpose, sequence coverage was defined as the average number of times that each base was independently read with a base-quality score corresponding to at least 99% accuracy.) We attained an overall average of 4.5-fold coverage across the genome for draft clones. A few of the sequenced clones fell below the minimum of threefold sequence coverage or have not been formally designated by centres as meeting draft standards; these are referred to as predraft (Table 2). Some of these are clones that span remaining gaps in the draft genome sequence and were in the process of being sequenced on 7 October 2000; a few are old submissions from centres that are no longer active.

The lengths of the initial sequence contigs in the draft clones vary as a function of coverage, but half of all nucleotides reside in initial sequence contigs of at least 21.7 kb (see below). Various properties of the draft clones can be assessed from instances in which there was substantial overlap between a draft clone and a finished (or nearly finished) clone. By examining the sequence alignments in the overlap regions, we estimated that the initial sequence contigs in a draft sequence clone cover an average of about 96% of the clone and are separated by gaps with an average size of about 500 bp.

Although the main emphasis was on producing a draft genome sequence, the centres also maintained sequence finishing activities during this period, leading to a twofold increase in finished sequence from June 1999 to June 2000 (Fig. 4). The total amount of human sequence in this final form stood at more than 835 Mb on 7 October 2000, or more than 25% of the human genome. This includes the finished sequences of chromosomes 21 and 22 (refs 93, 94). As centres have begun to shift from draft to finished sequencing in the last quarter of 2000, the production of finished sequence has increased to an annualized rate of 1 Gb per year and is continuing to rise.

Table 3 Total human sequence deposited in the HTGS division of GenBank

Sequencing centre	Total human sequence (kb)	Finished human sequence (kb)
Whitehead Institute, Center for Genome Research*	1,196,888	46,560
The Sanger Centre*	970,789	284,353
Washington University Genome Sequencing Center*	765,898	175,279
US DOE Joint Genome Institute	377,998	78,486
Baylor College of Medicine Human Genome Sequencing Center	345,125	53,418
RIKEN Genomic Sciences Center	203,166	16,971
Genoscope	85,995	48,808
GTC Sequencing Center	71,357	7,014
Department of Genome Analysis, Institute of Molecular Biotechnology	49,865	17,788
Beijing Genomics Institute/Human Genome Center	42,865	6,297
Multimegabase Sequencing Center; Institute for Systems Biology	31,241	9,676
Stanford Genome Technology Center	29,728	3,530
The Stanford Human Genome Center and Department of Genetics	28,162	9,121
University of Washington Genome Center	24,115	14,692
Keio University	17,364	13,058
University of Texas Southwestern Medical Center at Dallas	11,670	7,028
University of Oklahoma Advanced Center for Genome Technology	10,071	9,155
Max Planck Institute for Molecular Genetics	7,650	2,940
GBF – German Research Centre for Biotechnology	4,639	2,338
Cold Spring Harbor Laboratory Lita Annenberg Hazen Genome Center	4,338	2,104
Other	59,574	35,911
Total	4,338,224	842,027

Total human sequence deposited in GenBank by members of the International Human Genome Sequencing Consortium, as of 8 October 2000. The amount of total sequence (finished plus draft plus predraft) is shown in the second column and the amount of finished sequence is shown in the third column. Total sequence differs from totals in Tables 1 and 2 because of inclusion of padding characters and of some clones not used in assembly. HTGS, high throughput genome sequence.

*These three centres produced an additional 2.4 Gb of raw plasmid paired-end reads (see Table 4), consisting of 0.99 Gb from Whitehead Institute, 0.66 Gb from The Sanger Centre and 0.75 Gb from Washington University.

In addition to sequencing large-insert clones, three centres generated a large collection of random raw sequence reads from whole-genome shotgun libraries (Table 4; ref. 98). These 5.77 million successful sequences contained 2.4 Gb of high-quality bases; this corresponds to about 0.75-fold coverage and would be statistically expected to include about 50% of the nucleotides in the human genome (data available at <http://snp.cshl.org/data>). The primary objective of this work was to discover SNPs, by comparing these random raw sequences (which came from different individuals) with the draft genome sequence. However, many of these raw sequences were obtained from both ends of plasmid clones and thereby also provided valuable 'linking' information that was used in sequence assembly. In addition, the random raw sequences provide sequence coverage of about half of the nucleotides not yet represented in the sequenced large-insert clones; these can be used as probes for portions of the genome not yet recovered.

Assembly of the draft genome sequence

We then set out to assemble the sequences from the individual large-insert clones into an integrated draft sequence of the human genome. The assembly process had to resolve problems arising from the draft nature of much of the sequence, from the variety of clone sources, and from the high fraction of repeated sequences in the human genome. This process involved three steps: filtering, layout and merging.

The entire data set was filtered uniformly to eliminate contamination from nonhuman sequences and other artefacts that had not already been removed by the individual centres. (Information about contamination was also sent back to the centres, which are updating the individual entries in the public databases.) We also identified instances in which the sequence data from one BAC clone was substantially contaminated with sequence data from another (human or nonhuman) clone. The problems were resolved in most instances; 231 clones remained unresolved, and these were eliminated from the assembly reported here. Instances of lower levels of cross-contamination (for example, a single 96-well microplate misassigned to the wrong BAC) are more difficult to detect; some undoubtedly remain and may give rise to small spurious sequence contigs in the draft genome sequence. Such issues are readily resolved as the clones progress towards finished sequence, but they necessitate some caution in certain applications of the current data.

The sequenced clones were then associated with specific clones on the physical map to produce a 'layout'. In principle, sequenced clones that correspond to fingerprinted BACs could be directly assigned by name to fingerprint clone contigs on the fingerprint-based physical map. In practice, however, laboratory mixups occasionally resulted in incorrect assignments. To eliminate such problems, sequenced clones were associated with the fingerprint clone contigs in the physical map by using the sequence data to calculate a

Table 4 Plasmid paired-end reads

	Total reads deposited*	Read pairs†	Size range of inserts (kb)
Random-sheared	3,227,685	1,155,284	1.8–6
Enzyme digest	2,539,222	761,010	0.8–4.7
Total	5,766,907	1,916,294	

The plasmid paired-end reads used a mixture of DNA from a set of 24 samples from the DNA Polymorphism Discovery Resource (<http://locus.umdnj.edu/ngms/pdr.html>). This set of 24 anonymous US residents contains samples from European-Americans, African-Americans, Mexican-Americans, Native Americans and Asian-Americans, although the ethnicities of the individual samples are not identified. Informed consent to contribute samples to the DNA Polymorphism Discovery Resource was obtained from all 450 individuals who contributed samples. Samples from the European-American, African-American and Mexican-American individuals came from NHANES (<http://www.cdc.gov/nchs/nhanes.htm>); individuals were recontacted to obtain their consent for the Resource project. New samples were obtained from Asian-Americans whose ancestry was from a variety of East and South Asian countries. New samples were also obtained for the Native Americans; tribal permission was obtained first, and then individual consents. See http://www.nhgri.nih.gov/Grant_Info/Funding/RFA/discover_polymorphisms.html and ref. 98.

*Reflects data deposited with and released by The SNP Consortium (see <http://snp.cshl.org/data>). †Read pairs represents the number of cases in which sequence from both ends of a genomic cloned fragment was determined and used in this study as linking information.

partial list of restriction fragments *in silico* and comparing that list with the experimental database of BAC fingerprints. The comparison was feasible because the experimental sizing of restriction fragments was highly accurate (to within 0.5–1.5% of the true size, for 95% of fragments from 600 to 12,000 base pairs (bp))^{84,85}. Reliable matching scores could be obtained for 16,193 of the clones. The remaining sequenced clones could not be placed on the map by this method because they were too short, or they contained too many small initial sequence contigs to yield enough restriction fragments, or possibly because their sequences were not represented in the fingerprint database.

An independent approach to placing sequenced clones on the physical map used the database of end sequences from fingerprinted BACs (Table 1). Sequenced clones could typically be reliably mapped if they contained multiple matches to BAC ends, with all corresponding to clones from a single genomic region (multiple matches were required as a safeguard against errors known to exist in the BAC end database and against repeated sequences). This approach provided useful placement information for 22,566 sequenced clones.

Altogether, we could assign 25,403 sequenced clones to fingerprint clone contigs by combining *in silico* digestion and BAC end sequence match data. To place most of the remaining sequenced clones, we exploited information about sequence overlap or BAC-end paired links of these clones with already positioned clones. This left only a few, mostly small, sequenced clones that could not be placed (152 sequenced clones containing 5.5 Mb of sequence out of 29,298 sequenced clones containing more than 4,260 Mb of sequence); these are being localized by radiation hybrid mapping of STSs derived from their sequences.

The fingerprint clone contigs were then mapped to chromosomal locations, using sequence matches to mapped STSs from four human radiation hybrid maps^{95,99,100}, one YAC and radiation hybrid map²⁹, and two genetic maps^{101,102}, together with data from FISH^{86,90,103}. The mapping was iteratively refined by comparing the order and orientation of the STSs in the fingerprint clone contigs and the various STS-based maps, to identify and refine discrepancies (Fig. 5). Small fingerprint clone contigs (< 1 Mb) were difficult to orient and, sometimes, to order using these methods. In all, 942 fingerprint clone contigs contained sequenced clones. (An additional 304 of the 1,246 fingerprint clone contigs did not contain sequenced clones, but these tended to be extremely small and together contain less than 1% of the mapped clones. About one-third have been targeted for sequencing. A few derive from the Y chromosome, for which the map was constructed separately⁸⁹. Most of the remainder are fragments of other larger contigs or represent other artefacts. These are being eliminated in subsequent versions of the database.) Of these 942 contigs with sequenced clones, 852 (90%, containing 99.2% of the total sequence) were localized to specific chromosome locations in this way. An additional 51 fingerprint clone contigs, containing 0.5% of the sequence, could be assigned to a specific chromosome but not to a precise position.

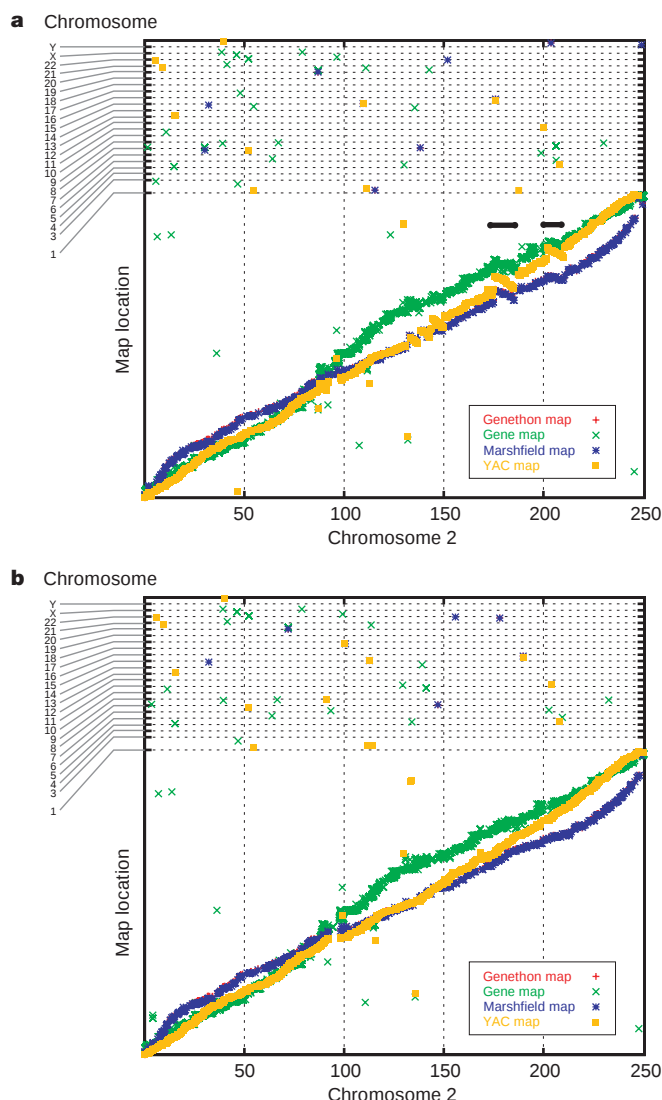


Figure 5 Positions of markers on previous maps of the genome (the Genethon¹⁰¹ genetic map and Marshfield genetic map (http://research.marshfieldclinic.org/genetics/genotyping_service/mgsver2.htm), the GeneMap99 radiation hybrid map¹⁰⁰, and the Whitehead YAC and radiation hybrid map²⁹) plotted against their derived position on the draft sequence for chromosome 2. The horizontal units are Mb but the vertical units of each map vary (cM, cR and so on) and thus all were scaled so that the entire map spans the full vertical range. Markers that map to other chromosomes are shown in the chromosome lines at the top. The data sets generally follow the diagonal, indicating that order and orientation of the marker sets on the different maps largely agree (note that the two genetic maps are completely superimposed). In **a**, there are two segments (bars) that are inverted in an earlier version draft sequence relative to all the other maps. **b**, The same chromosome after the information was used to reorient those two segments.

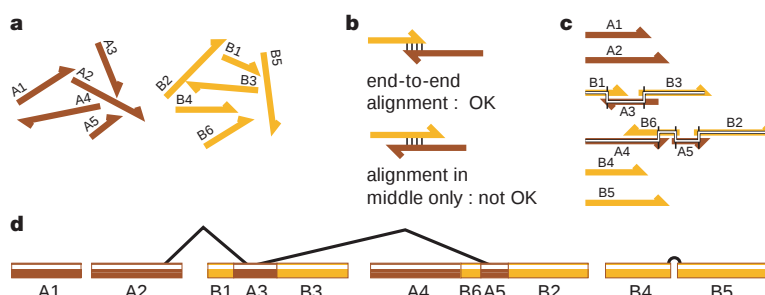


Figure 6 The key steps (**a–d**) in assembling individual sequenced clones into the draft genome sequence. A1–A5 represent initial sequence contigs derived from shotgun sequencing of clone A, and B1–B6 are from clone B.

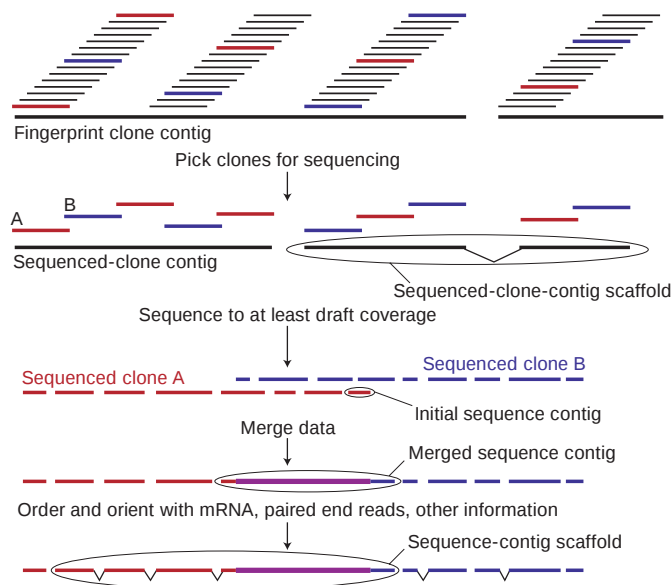


Figure 7 Levels of clone and sequence coverage. A 'fingerprint clone contig' is assembled by using the computer program FPC^{84,451} to analyse the restriction enzyme digestion patterns of many large-insert clones. Clones are then selected for sequencing to minimize overlap between adjacent clones. For a clone to be selected, all of its restriction enzyme fragments (except the two vector-insert junction fragments) must be shared with at least one of its neighbours on each side in the contig. Once these overlapping clones have been sequenced, the set is a 'sequenced-clone contig'. When all selected clones from a fingerprint clone contig have been sequenced, the sequenced-clone contig will be the same as the fingerprint clone contig. Until then, a fingerprint clone contig may contain several sequenced-clone contigs. After individual clones (for example, A and B) have been sequenced to draft coverage and the clones have been mapped, the data are analysed by GigAssembler (Fig. 6), producing merged sequence contigs from initial sequence contigs, and linking these to form sequence-contig scaffolds (see Box 1).

The remaining 39 contigs containing 0.3% of the sequence were not positioned at all.

We then merged the sequences from overlapping sequenced clones (Fig. 6), using the computer program GigAssembler¹⁰⁴. The program considers nearby sequenced clones, detects overlaps between the initial sequence contigs in these clones, merges the overlapping sequences and attempts to order and orient the sequence contigs. It begins by aligning the initial sequence contigs from one clone with those from other clones in the same fingerprint clone contig on the basis of length of alignment, per cent identity of the alignment, position in the sequenced clone layout and other factors. Alignments are limited to one end of each initial sequence contig for partially overlapping contigs or to both ends of an initial sequence contig contained entirely within another; this eliminates internal alignments that may reflect repeated sequence or possible misassembly (Fig. 6b). Beginning with the highest scoring pairs, initial sequence contigs are then integrated to produce 'merged sequence contigs' (usually referred to simply as 'sequence contigs'). The program refines the arrangement of the clones within the fingerprint clone contig on the basis of the extent of sequence overlap between them and then rebuilds the sequence contigs. Next, the program selects a sequence path through the sequence contigs (Fig. 6c). It tries to use the highest quality data by preferring longer initial sequence contigs and avoiding the first and last 250 bases of initial sequence contigs where possible. Finally, it attempts to order and orient the sequence contigs by using additional information, including sequence data from paired-end plasmid and BAC reads, known messenger RNAs and ESTs, as well as additional linking information provided by centres. The sequence contigs are thereby linked together to create 'sequence-contig scaffolds' (Fig. 6d). The process also joins overlapping sequenced clones into sequenced-clone contigs and links sequenced-clone contigs to form sequenced-clone-contig scaffolds. A fingerprint clone contig may contain several sequenced-clone contigs, because bridging clones remain to be sequenced. The assembly contained 4,884 sequenced-clone

Table 5 The draft genome sequence

Chromosome	Sequence from clones (kb)			Sequence from contigs (kb)		
	Finished clones	Draft clones	Pre-draft clones	Contigs containing finished clones	Deep coverage sequence contigs	Draft/pre-draft sequence contigs
All	826,441	1,734,995	131,476	958,922	840,815	893,175
1	50,851	149,027	12,356	61,001	78,773	72,461
2	46,909	167,439	7,210	53,775	81,569	86,214
3	22,350	152,840	11,057	26,959	79,649	79,638
4	15,914	134,973	17,261	19,096	66,165	82,887
5	37,973	129,581	2,160	48,895	61,387	59,431
6	75,312	76,082	6,696	93,458	28,204	36,428
7	94,845	47,328	4,047	103,188	14,434	28,597
8	14,538	102,484	7,236	16,659	47,198	60,400
9	18,401	77,648	10,864	24,030	42,653	40,230
10	16,889	99,181	11,066	21,421	54,054	51,662
11	13,162	111,092	4,352	16,145	65,147	47,314
12	32,156	84,653	7,651	37,519	43,995	42,946
13	16,818	68,983	7,136	22,191	38,319	32,429
14	58,989	27,370	565	78,302	3,267	5,355
15	2,739	67,453	3,211	3,112	34,758	35,533
16	22,987	48,997	1,143	27,751	20,892	24,484
17	29,881	36,349	6,600	33,531	14,671	24,628
18	5,128	65,284	2,352	6,656	40,947	25,160
19	28,481	26,568	369	32,228	7,188	16,003
20	54,217	5,302	976	56,534	1,065	2,896
21	33,824	0	0	33,824	0	0
22	33,786	0	0	33,786	0	0
X	77,630	45,100	4,941	83,796	14,056	29,820
Y	18,169	3,221	363	20,222	333	1,198
NA	2,434	1,858	844	2,446	122	2,568
UL	2,056	6,182	1,020	2,395	1,969	4,894

The table presents summary statistics for the draft genome sequence over the entire genome and by individual chromosome. NA, clones that could not be placed into the sequenced clone layout. UL, clones that could be placed in the layout, but that could not reliably be placed on a chromosome. First three columns, data from finished clones, draft clones and pre-draft clones. The last three columns break the data down according to the type of sequence contig. Contigs containing finished clones represent sequence contigs that consist of finished sequence plus any (small) extensions from merged sequence contigs that arise from overlap with flanking draft clones. Deep coverage sequence contigs include sequence from two or more overlapping unfinished clones; they consist of roughly full shotgun coverage and thus are longer than the average unfinished sequence contig. Draft/pre-draft sequence contigs are all of the other sequence contigs in unfinished clones. Thus, the draft genome sequence consists of approximately one-third finished sequence, one-third deep coverage sequence and one-third draft/pre-draft coverage sequence. In all of the statistics, we count only nonoverlapping bases in the draft genome sequence.

contigs in 942 fingerprint clone contigs.

The hierarchy of contigs is summarized in Fig. 7. Initial sequence contigs are integrated to create merged sequence contigs, which are then linked to form sequence-contig scaffolds. These scaffolds reside within sequenced-clone contigs, which in turn reside within fingerprint clone contigs.

The draft genome sequence

The result of the assembly process is an integrated draft sequence of the human genome. Several features of the draft genome sequence are reported in Tables 5–7, including the proportion represented by finished, draft and predraft categories. The Tables also show the numbers and lengths of different types of contig, for each chromosome and for the genome as a whole.

The contiguity of the draft genome sequence at each level is an important feature. Two commonly used statistics have significant drawbacks for describing contiguity. The 'average length' of a contig is deflated by the presence of many small contigs comprising only a small proportion of the genome, whereas the 'length-weighted average length' is inflated by the presence of large segments of finished sequence. Instead, we chose to describe the contiguity as a property of the 'typical' nucleotide. We used a statistic called the 'N50 length', defined as the largest length L such that 50% of all nucleotides are contained in contigs of size at least L .

The continuity of the draft genome sequence reported here and the effectiveness of assembly can be readily seen from the following: half of all nucleotides reside within an initial sequence contig of at least 21.7 kb, a sequence contig of at least 82 kb, a sequence-contig scaffold of at least 274 kb, a sequenced-clone contig of at least 826 kb and a fingerprint clone contig of at least 8.4 Mb (Tables 6, 7). The cumulative distributions for each of these measures of contiguity are shown in Fig. 8, in which the N50 values for each measure can be seen as the value at which the cumulative distributions cross 50%. We have also estimated the size of each chromosome, by estimating the gap sizes (see below) and the extent of missing heterochromatic sequence^{93,94,105–108} (Table 8). This is undoubtedly an oversimplification and does not adequately take into account the sequence status of each chromosome. Nonetheless, it provides a useful way to relate the draft sequence to the chromosomes.

Quality assessment

The draft genome sequence already covers the vast majority of the genome, but it remains an incomplete, intermediate product that is regularly updated as we work towards a complete finished sequence. The current version contains many gaps and errors. We therefore sought to evaluate the quality of various aspects of the current draft genome sequence, including the sequenced clones themselves, their assignment to a position in the fingerprint clone contigs, and the assembly of initial sequence contigs from the individual clones into sequence-contig scaffolds.

Nucleotide accuracy is reflected in a PHRAP score assigned to each base in the draft genome sequence and available to users through the Genome Browsers (see below) and public database entries. A summary of these scores for the unfinished portion of the genome is shown in Table 9. About 91% of the unfinished draft genome sequence has an error rate of less than 1 per 10,000 bases (PHRAP score > 40), and about 96% has an error rate of less than 1 in 1,000 bases (PHRAP > 30). These values are based only on the quality scores for the bases in the sequenced clones; they do not reflect additional confidence in the sequences that are represented in overlapping clones. The finished portion of the draft genome sequence has an error rate of less than 1 per 10,000 bases.

Individual sequenced clones. We assessed the frequency of misassemblies, which can occur when the assembly program PHRAP joins two nonadjacent regions in the clone into a single initial sequence contig. The frequency of misassemblies depends heavily on the depth and quality of coverage of each clone and the nature of the underlying sequence; thus it may vary among genomic regions and among individual centres. Most clone misassemblies are readily corrected as coverage is added during finishing, but they may have been propagated into the current version of the draft genome sequence and they justify caution for certain applications.

We estimated the frequency of misassembly by examining instances in which there was substantial overlap between a draft clone and a finished clone. We studied 83 Mb of such overlaps, involving about 9,000 initial sequence contigs. We found 5.3 instances per Mb in which the alignment of an initial sequence contig to the finished sequence failed to extend to within 200 bases

Table 6 Clone level contiguity of the draft genome sequence

Chromosome	Sequenced-clone contigs		Sequenced-clone-contig scaffolds		Fingerprint clone contigs with sequence	
	Number	N50 length (kb)	Number	N50 length (kb)	Number	N50 length (kb)
All	4,884	826	2,191	2,279	942	8,398
1	453	650	197	1,915	106	3,537
2	348	1,028	127	3,140	52	10,628
3	409	672	201	1,550	73	5,077
4	384	606	163	1,659	41	6,918
5	385	623	164	1,642	48	5,747
6	292	814	98	3,292	17	24,680
7	224	1,074	86	3,527	29	20,401
8	292	542	115	1,742	43	6,236
9	143	1,242	78	2,411	21	29,108
10	179	1,097	105	1,952	16	30,284
11	224	887	89	3,024	31	9,414
12	196	1,138	76	2,717	28	9,546
13	128	1,151	56	3,257	13	25,256
14	54	3,079	27	8,489	14	22,128
15	123	797	56	2,095	19	8,274
16	159	620	92	1,317	57	2,716
17	138	831	58	2,138	43	2,816
18	137	709	47	2,572	24	4,887
19	159	569	79	1,200	51	1,534
20	42	2,318	20	6,862	9	23,489
21	5	28,515	5	28,515	5	28,515
22	11	23,048	11	23,048	11	23,048
X	325	572	181	1,082	143	1,436
Y	27	1,539	20	3,290	8	5,135
UL	47	227	40	281	40	281

Number and size of sequenced-clone contigs, sequenced-clone-contig scaffolds and those fingerprint clone contigs (see Box 1) that contain sequenced clones; some small fingerprint clone contigs do not as yet have associated sequence. UL, fingerprint clone contigs that could not reliably be placed on a chromosome. These length estimates are from the draft genome sequence, in which gaps between sequence contigs are arbitrarily represented with 100 Ns and gaps between sequence clone contigs with 50,000 Ns for 'bridged gaps' and 100,000 Ns for 'unbridged gaps'. These arbitrary values differ minimally from empirical estimates of gap size (see text), and using the empirically derived estimates would change the N50 lengths presented here only slightly. For unfinished chromosomes, the N50 length ranges from 1.5 to 3 times the arithmetic mean for sequenced-clone contigs, 1.5 to 3 times for sequenced-clone-contig scaffolds, and 1.5 to 6 times for fingerprint clone contigs with sequence.

of the end of the contig, suggesting a possible false join in the assembly of the initial sequence contig. In about half of these cases, the potential misassembly involved fewer than 400 bases, suggesting that a single raw sequence read may have been incorrectly joined. We found 1.9 instances per Mb in which the alignment showed an internal gap, again suggesting a possible misassembly; and 0.5 instances per Mb in which the alignment indicated that two initial sequence contigs that overlapped by at least 150 bp had not been merged by PHRAP. Finally, there were another 0.9 instances per Mb with various other problems. This gives a total of 8.6 instances per Mb of possible misassembly, with about half being relatively small issues involving a few hundred bases.

Some of the potential problems might not result from misassembly, but might reflect sequence polymorphism in the population,

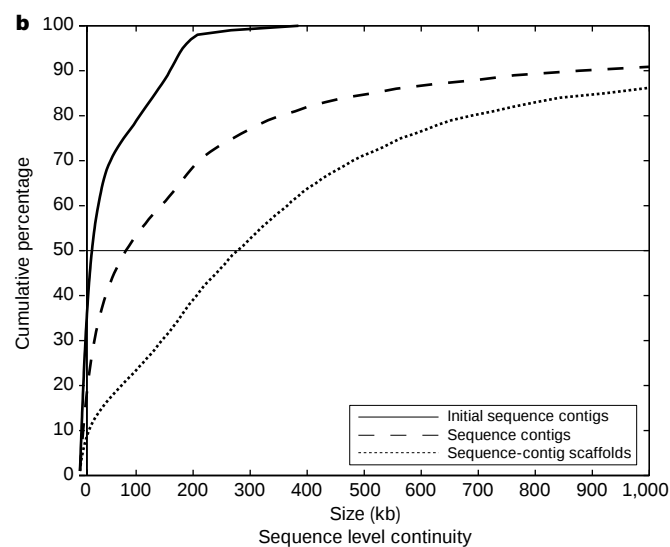
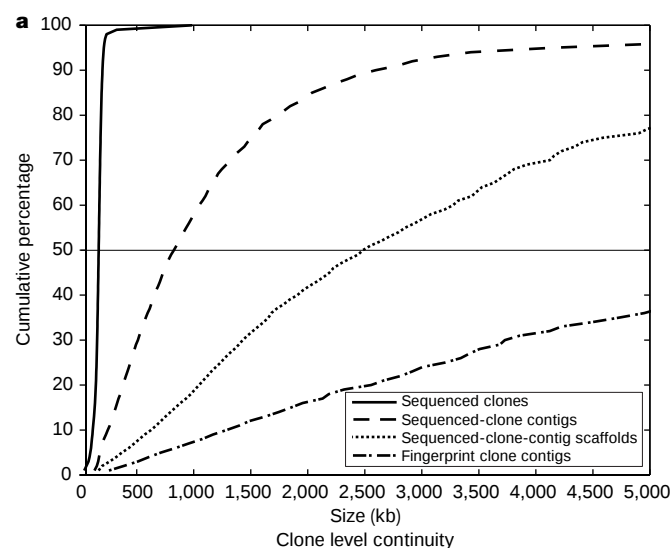


Figure 8 Cumulative distributions of several measures of clone level continuity and sequence continuity. The figures represent the proportion of the draft genome sequence contained in contigs of at most the indicated size. **a**, Clone level continuity. The clones have a tight size distribution with an N50 of ~160 kb (corresponding to 50% on the cumulative distribution). Sequenced-clone contigs represent the next level of continuity, and are linked by mRNA sequences or pairs of BAC end sequences to yield the sequenced-clone-contig scaffolds. The underlying continuity of the layout of sequenced clones against the fingerprinted clone contigs is only partially shown at this scale. **b**, Sequence continuity. The input fragments have low continuity (N50 = 21.7 kb). After merging, the sequence contigs grow to an N50 length of about 82 kb. After linking, sequence-contig scaffolds with an N50 length of about 274 kb are created.

small rearrangements during growth of the large-insert clones, regions of low-quality sequence or matches between segmental duplications. Thus, the frequency of misassemblies may be overstated. On the other hand, the criteria for recognizing overlap between draft and finished clones may have eliminated some misassemblies.

Layout of the sequenced clones. We assessed the accuracy of the layout of sequenced clones onto the fingerprinted clone contigs by calculating the concordance between the positions assigned to a sequenced clone on the basis of *in silico* digestion and the position assigned on the basis of BAC end sequence data. The positions agreed in 98% of cases in which independent assignments could be made by both methods. The results were also compared with well studied regions containing both finished and draft genome sequence. These results indicated that sequenced clone order in the fingerprint map was reliable to within about half of one clone length (~100 kb).

A direct test of the layout is also provided by the draft genome sequence assembly itself. With extensive coverage of the genome, a correctly placed clone should usually (although not always) show sequence overlap with its neighbours in the map. We found only 421 instances of 'singleton' clones that failed to overlap a neighbouring clone. Close examination of the data suggests that most of these are correctly placed, but simply do not yet overlap an adjacent sequenced clone. About 150 clones appeared to be candidates for being incorrectly placed.

Alignment of the fingerprint clone contigs. The alignment of the fingerprint clone contigs with the chromosomes was based on the radiation hybrid, YAC and genetic maps of STSs. The positions of most of the STSs in the draft genome sequence were consistent with these previous maps, but the positions of about 1.7% differed from one or more of them. Some of these disagreements may be due to errors in the layout of the sequenced clones or in the underlying

Figure 9 Overview of features of draft human genome. The Figure shows the occurrences of twelve important types of feature across the human genome. Large grey blocks represent centromeres and centromeric heterochromatin (size not precisely to scale). Each of the feature types is depicted in a track, from top to bottom as follows. (1) Chromosome position in Mb. (2) The approximate positions of Giemsa-stained chromosome bands at the 800 band resolution. (3) Level of coverage in the draft genome sequence. Red, areas covered by finished clones; yellow, areas covered by predraft sequence. Regions covered by draft sequenced clones are in orange, with darker shades reflecting increasing shotgun sequence coverage. (4) GC content. Percentage of bases in a 20,000 base window that are C or G. (5) Repeat density. Red line, density of SINE class repeats in a 100,000-base window; blue line, density of LINE class repeats in a 100,000-base window. (6) Density of SNPs in a 50,000-base window. The SNPs were detected by sequencing and alignments of random genomic reads. Some of the heterogeneity in SNP density reflects the methods used for SNP discovery. Rigorous analysis of SNP density requires comparing the number of SNPs identified to the precise number of bases surveyed. (7) Non-coding RNA genes. Brown, functional RNA genes such as tRNAs, snoRNAs and rRNAs; light orange, RNA pseudogenes. (8) CpG islands. Green ticks represent regions of ~200 bases with CpG levels significantly higher than in the genome as a whole, and GC ratios of at least 50%. (9) Exonfish ecores. Regions of homology with the pufferfish *T. nigroviridis*²⁹² are blue. (10) ESTs with at least one intron when aligned against genomic DNA are shown as black tick marks. (11) The starts of genes predicted by Genie or Ensembl are shown as red ticks. The starts of known genes from the RefSeq database¹¹⁰ are shown in blue. (12) The names of genes that have been uniquely located in the draft genome sequence, characterized and named by the HGM Nomenclature Committee. Known disease genes from the OMIM database are red, other genes blue. This Figure is based on an earlier version of the draft genome sequence than analysed in the text, owing to production constraints. We are aware of various errors in the Figure, including omissions of some known genes and misplacements of others. Some genes are mapped to more than one location, owing to errors in assembly, close paralogues or pseudogenes. Manual review was performed to select the most likely location in these cases and to correct other regions. For updated information, see <http://genome.ucsc.edu/> and <http://www.ensembl.org/>.

Table 7 Sequence level contiguity of the draft genome sequence

Chromosome	Initial sequence contigs		Sequence contigs		Sequence-contig scaffolds	
	Number	N50 length (kb)	Number	N50 length (kb)	Number	N50 length (kb)
All	396,913	21.7	149,821	81.9	87,757	274.3
1	37,656	16.5	12,256	59.1	5,457	278.4
2	32,280	19.9	13,228	57.3	6,959	248.5
3	38,848	15.6	15,098	37.7	8,964	167.4
4	28,600	16.0	13,152	33.0	7,402	158.9
5	30,096	20.4	10,689	72.9	6,378	241.2
6	17,472	43.6	5,547	180.3	2,554	485.0
7	12,733	86.4	4,562	335.7	2,726	591.3
8	19,042	18.1	8,984	38.2	4,631	198.9
9	15,955	20.1	6,226	55.6	3,766	216.2
10	21,762	18.7	9,126	47.9	6,886	133.0
11	29,723	14.3	8,503	40.0	4,684	193.2
12	22,050	19.1	8,422	63.4	5,526	217.0
13	13,737	21.7	5,193	70.5	2,659	300.1
14	4,470	161.4	829	1,371.0	541	2,009.5
15	13,134	15.3	5,840	30.3	3,229	149.7
16	10,297	34.4	4,916	119.5	3,337	356.3
17	10,369	22.9	4,339	90.6	2,616	248.9
18	16,266	15.3	4,461	51.4	2,540	216.1
19	6,009	38.4	2,503	134.4	1,551	375.5
20	2,884	108.6	511	1,346.7	312	813.8
21	103	340.0	5	28,515.3	5	28,515.3
22	526	113.9	11	23,048.1	11	23,048.1
X	11,062	58.8	4,607	218.6	2,610	450.7
Y	557	154.3	140	1,388.6	106	1,439.7
UL	1,282	21.4	613	46.0	297	166.4

This Table is similar to Table 6 but shows the number and N50 length for various types of sequence contig (see Box 1). See legend to Table 6 concerning treatment of gaps. For sequence contigs in the draft genome sequence, the N50 length ranges from 1.7 to 5.5 times the arithmetic mean for initial sequence contigs, 2.5 to 8.2 times for merged sequence contigs, and 6.1 to 10 times for sequence-contig scaffolds.

Table 8 Chromosome size estimates

Chromosome*	Sequenced bases† (Mb)	FCC gaps‡		SCC gaps§		Sequence gaps#		Heterochromatin and short arm adjustments** (Mb)	Total estimated chromosome size (including artefactual duplication in draft genome sequence)†† (Mb)	Previously estimated chromosome size‡‡ (Mb)
		Number	Total bases in gaps§ (Mb)	Number	Total bases in gaps¶ (Mb)	Number	Total bases in gaps# (Mb)			
All	2,692.9	897	152.0	4,076	142.7	145,514	80.6	212	3,289	3,286
1	212.2	104	17.7	347	12.1	11,803	6.5	30	279	263
2	221.6	50	8.5	296	10.4	12,880	7.1	3	251	255
3	186.2	71	12.1	336	11.8	14,689	8.1	3	221	214
4	168.1	39	6.6	343	12.0	12,768	7.1	3	197	203
5	169.7	46	7.8	337	11.8	10,304	5.7	3	198	194
6	158.1	15	2.6	275	9.6	5,225	2.9	3	176	183
7	146.2	27	4.6	195	6.8	4,338	2.4	3	163	171
8	124.3	41	7.0	249	8.7	8,692	4.8	3	148	155
9	106.9	19	3.2	122	4.3	6,083	3.4	22	140	145
10	127.1	14	2.4	163	5.7	8,947	5.0	3	143	144
11	128.6	29	4.9	193	6.8	8,279	4.6	3	148	144
12	124.5	26	4.4	168	5.9	8,226	4.6	3	142	143
13	92.9	12	2.0	115	4.0	5,065	2.8	16	118	114
14	86.9	13	2.2	40	1.4	775	0.4	16	107	109
15	73.4	18	3.1	104	3.6	5,717	3.2	17	100	106
16	73.1	55	9.4	102	3.6	4,757	2.6	15	104	98
17	72.8	41	7.0	95	3.3	4,261	2.4	3	88	92
18	72.9	22	3.7	113	4.0	4,324	2.4	3	86	85
19	55.4	49	8.3	108	3.8	2,344	1.3	3	72	67
20	60.5	7	1.2	33	1.2	469	0.3	3	66	72
21	33.8	4	0.1	0	0.0	0	0.0	11	45	50
22	33.8	10	1.0	0	0.0	0	0.0	13	48	56
X	127.7	141	24.0	182	6.4	4,282	2.4	3	163	164
Y	21.8	6	1.0	19	0.7	113	0.1	27	51	59
NA	5.1	0	0	134	0.0	577	0.3	0	0	0
UL	9.3	38	0	7	0.0	566	0.3	0	0	0

* NA, sequenced clones that could not be associated with fingerprint clone contigs. UL, clone contigs that could not be reliably placed on a chromosome.

† Total number of bases in the draft genome sequence, excluding gaps. Total length of scaffold (including gaps contained within clones) is 2.916 Gb.

‡ Gaps between those fingerprint clone contigs that contain sequenced clones excluding gaps for centromeres.

§ For unfinished chromosomes, we estimate an average size of 0.17 Mb per FCC gap, based on retrospective estimates of the clone coverage of chromosomes 21 and 22. Gap estimates for chromosomes 21 and 22 are taken from refs 93, 94.

¶ Gaps between sequenced-clone contigs within a fingerprint clone contig.

‡‡ For unfinished chromosomes, we estimate sequenced clone gaps at 0.035 Mb each, based on evaluation of a sample of these gaps.

Gaps between two sequence contigs within a sequenced-clone contig.

** We estimate the average number of bases in sequence gaps from alignments of the initial sequence contigs of unfinished clones (see text) and extrapolation to the whole chromosome.

†† Including adjustments for estimates of the sizes of the short arms of the acrocentric chromosomes 13, 14, 15, 21 and 22 (ref. 105), estimates for the centromere and heterochromatic regions of chromosomes 1, 9 and 16 (refs 106, 107) and estimates of 3 Mb for the centromere and 24 Mb for telomeric heterochromatin for the Y chromosome¹⁰⁸.

‡‡ The sum of the five lengths in the preceding columns. This is an overestimate, because the draft genome sequence contains some artefactual sequence owing to inability to correctly merge all underlying sequence contigs. The total amount of artefactual duplication varies among chromosomes; the overall amount is estimated by computational analysis to be about 100 Mb, or about 3% of the total length given, yielding a total estimated size of about 3,200 Mb for the human genome.

‡‡ Including heterochromatic regions and acrocentric short arm(s)¹⁰⁵.

fingerprint map. However, many involve STSs that have been localized on only one or two of the previous maps or that occur as isolated discrepancies in conflict with several flanking STSs. Many of these cases are probably due to errors in the previous maps (with error rates for individual maps estimated at 1–2%¹⁰⁰). Others may be due to incorrect assignment of the STSs to the draft genome sequence (by the electronic polymerase chain reaction (e-PCR) computer program) or to database entries that contain sequence data from more than one clone (owing to cross-contamination).

Graphical views of the independent data sets were particularly useful in detecting problems with order or orientation (Fig. 5). Areas of conflict were reviewed and corrected if supported by the underlying data. In the version discussed here, there were 41 sequenced clones falling in 14 sequenced-clone contigs with STS content information from multiple maps that disagreed with the flanking clones or sequenced-clone contigs; the placement of these clones thus remains suspect. Four of these instances suggest errors in the fingerprint map, whereas the others suggest errors in the layout of sequenced clones. These cases are being investigated and will be corrected in future versions.

Assembly of the sequenced clones. We assessed the accuracy of the assembly by using a set of 148 draft clones comprising 22.4 Mb for which finished sequence subsequently became available¹⁰⁴. The initial sequence contigs lack information about order and orientation, and GigAssembler attempts to use linking data to infer such information as far as possible¹⁰⁴. Starting with initial sequence contigs that were unordered and unoriented, the program placed 90% of the initial sequence contigs in the correct orientation and 85% in the correct order with respect to one another. In a separate test, GigAssembler was tested on simulated draft data produced from finished sequence on chromosome 22 and similar results were obtained.

Some problems remain at all levels. First, errors in the initial sequence contigs persist in the merged sequence contigs built from them and can cause difficulties in the assembly of the draft genome sequence. Second, GigAssembler may fail to merge some overlapping sequences because of poor data quality, allelic differences or misassemblies of the initial sequence contigs; this may result in apparent local duplication of a sequence. We have estimated by various methods the amount of such artefactual duplication in the assembly from these and other sources to be about 100 Mb. On the other hand, nearby duplicated sequences may occasionally be incorrectly merged. Some sequenced clones remain incorrectly placed on the layout, as discussed above, and others (< 0.5%) remain unplaced. The fingerprint map has undoubtedly failed to resolve some closely related duplicated regions, such as the Williams region and several highly repetitive subtelomeric and pericentric regions (see below). Detailed examination and sequence finishing may be required to sort out these regions precisely, as has been done with chromosome Y⁸⁹. Finally, small sequenced-clone contigs with limited or no STS

landmark content remain difficult to place. Full utilization of the higher resolution radiation hybrid map (the TNG map) may help in this⁹⁵. Future targeted FISH experiments and increased map continuity will also facilitate positioning of these sequences.

Genome coverage

We next assessed the nature of the gaps within the draft genome sequence, and attempted to estimate the fraction of the human genome not represented within the current version.

Gaps in draft genome sequence coverage. There are three types of gap in the draft genome sequence: gaps within unfinished sequenced clones; gaps between sequenced-clone contigs, but within fingerprint clone contigs; and gaps between fingerprint clone contigs. The first two types are relatively straightforward to close simply by performing additional sequencing and finishing on already identified clones. Closing the third type may require screening of additional large-insert clone libraries and possibly new technologies for the most recalcitrant regions. We consider these three cases in turn.

We estimated the size of gaps within draft clones by studying instances in which there was substantial overlap between a draft clone and a finished clone, as described above. The average gap size in these draft sequenced clones was 554 bp, although the precise estimate was sensitive to certain assumptions in the analysis. Assuming that the sequence gaps in the draft genome sequence are fairly represented by this sample, about 80 Mb or about 3% (likely range 2–4%) of sequence may lie in the 145,514 gaps within draft sequenced clones.

The gaps between sequenced-clone contigs but within fingerprint clone contigs are more difficult to evaluate directly, because the draft genome sequence flanking many of the gaps is often not precisely aligned with the fingerprinted clones. However, most are much smaller than a single BAC. In fact, nearly three-quarters of these gaps are bridged by one or more individual BACs, as indicated by linking information from BAC end sequences. We measured the sizes of a subset of gaps directly by examining restriction fragment fingerprints of overlapping clones. A study of 157 'bridged' gaps and 55 'unbridged' gaps gave an average gap size of 25 kb. Allowing for the possibility that these gaps may not be fully representative and that some restriction fragments are not included in the calculation, a more conservative estimate of gap size would be 35 kb. This would indicate that about 150 Mb or 5% of the human genome may reside in the 4,076 gaps between sequenced-clone contigs. This sequence should be readily obtained as the clones spanning them are sequenced.

The size of the gaps between fingerprint clone contigs was estimated by comparing the fingerprint maps to the essentially completed chromosomes 21 and 22. The analysis shows that the fingerprinted BAC clones in the global database cover 97–98% of the sequenced portions of those chromosomes⁸⁶. The published sequences of these chromosomes also contain a few small gaps (5 and 11, respectively) amounting to some 1.6% of the euchromatic sequence, and do not include the heterochromatic portion. This suggests that the gaps between contigs in the fingerprint map contain about 4% of the euchromatic genome. Experience with closure of such gaps on chromosomes 20 and 7 suggests that many of these gaps are less than one clone in length and will be closed by clones from other libraries. However, recovery of sequence from these gaps represents the most challenging aspect of producing a complete finished sequence of the human genome.

As another measure of the representation of the BAC libraries, Riethman¹⁰⁹ has found BAC or cosmid clones that link to telomeric half-YACs or to the telomeric sequence itself for 40 of the 41 non-satellite telomeres. Thus, the fingerprint map appears to have no substantial gaps in these regions. Many of the pericentric regions are also represented, but analysis is less complete here (see below).

Representation of random raw sequences. In another approach to measuring coverage, we compared a collection of random raw sequence reads to the existing draft genome sequence. In principle,

Table 9 Distribution of PHRAP scores in the draft genome sequence

PHRAP score	Percentage of bases in the draft genome sequence
0–9	0.6
10–19	1.3
20–29	2.2
30–39	4.8
40–49	8.1
50–59	8.7
60–69	9.0
70–79	12.1
80–89	17.3
>90	35.9

PHRAP scores are a logarithmically based representation of the error probability. A PHRAP score of X corresponds to an error probability of $10^{-X/10}$. Thus, PHRAP scores of 20, 30 and 40 correspond to accuracy of 99%, 99.9% and 99.99%, respectively. PHRAP scores are derived from quality scores of the underlying sequence reads used in sequence assembly. See <http://www.genome.washington.edu/UWGC/analysis/analysis/phrap.htm>.

the fraction of reads matching the draft genome sequence should provide an estimate of genome coverage. In practice, the comparison is complicated by the need to allow for repeat sequences, the imperfect sequence quality of both the raw sequence and the draft genome sequence, and the possibility of polymorphism. Nonetheless, the analysis provides a reasonable view of the extent to which the genome is represented in the draft genome sequence and the public databases.

We compared the raw sequence reads against both the sequences used in the construction of the draft genome sequence and all of GenBank using the BLAST computer program. Of the 5,615 raw sequence reads analysed (each containing at least 100 bp of contiguous non-repetitive sequence), 4,924 had a match of $\geq 97\%$ identity with a sequenced clone, indicating that $88 \pm 1.5\%$ of the genome was represented in sequenced clones. The estimate is subject to various uncertainties. Most serious is the proportion of repeat sequence in the remainder of the genome. If the unsequenced portion of the genome is unusually rich in repeated sequence, we would underestimate its size (although the excess would be comprised of repeated sequence).

We examined those raw sequences that failed to match by comparing them to the other publicly available sequence resources. Fifty (0.9%) had matches in public databases containing cDNA sequences, STSs and similar data. An additional 276 (or 43% of the remaining raw sequence) had matches to the whole-genome shotgun reads discussed above (consistent with the idea that these reads cover about half of the genome).

We also examined the extent of genome coverage by aligning the cDNA sequences for genes in the RefSeq dataset¹¹⁰ to the draft genome sequence. We found that 88% of the bases of these cDNAs could be aligned to the draft genome sequence at high stringency (at least 98% identity). (A few of the alignments with either the random raw sequence reads or the cDNAs may be to a highly similar region in the genome, but such matches should affect the estimate of genome coverage by considerably less than 1%, based on the estimated extent of duplication within the genome (see below).)

These results indicate that about 88% of the human genome is represented in the draft genome sequence and about 94% in the combined publicly available sequence databases. The figure of 88% agrees well with our independent estimates above that about 3%, 5% and 4% of the genome reside in the three types of gap in the draft genome sequence.

Finally, a small experimental check was performed by screening a large-insert clone library with probes corresponding to 16 of the whole genome shotgun reads that failed to match the draft genome sequence. Five hybridized to many clones from different fingerprint clone contigs and were discarded as being repetitive. Of the remaining eleven, two fell within sequenced clones (presumably within sequence gaps of the first type), eight fell in fingerprint clone

contigs but between sequenced clones (gaps of the second type) and one failed to identify clones in the fingerprint map (gaps of the third type) but did identify clones in another large-insert library. Although these numbers are small, they are consistent with the view that the much of the remaining genome sequence lies within already identified clones in the current map.

Estimates of genome and chromosome sizes. Informed by this analysis of genome coverage, we proceeded to estimate the sizes of the genome and each of the chromosomes (Table 8). Beginning with the current assigned sequence for each chromosome, we corrected for the known gaps on the basis of their estimated sizes (see above). We attempted to account for the sizes of centromeres and heterochromatin, neither of which are well represented in the draft sequence. Finally, we corrected for around 100 Mb of artefactual duplication in the assembly. We arrived at a total human genome size estimate of around 3,200 Mb, which compares favourably with previous estimates based on DNA content.

We also independently estimated the size of the euchromatic portion of the genome by determining the fraction of the 5,615 random raw sequences that matched the finished portion of the human genome (whose total length is known with greater precision). Twenty-nine per cent of these raw sequences found a match among 835 Mb of nonredundant finished sequence. This leads to an estimate of the euchromatic genome size of 2.9 Gb. This agrees reasonably with the prediction above based on the length of the draft genome sequence (Table 8).

Update. The results above reflect the data on 7 October 2000. New data are continually being added, with improvements being made to the physical map, new clones being sequenced to close gaps and draft clones progressing to full shotgun coverage and finishing. The draft genome sequence will be regularly reassembled and publicly released.

Currently, the physical map has been refined such that the number of fingerprint clone contigs has fallen from 1,246 to 965; this reflects the elimination of some artefactual contigs and the closure of some gaps. The sequence coverage has risen such that 90% of the human genome is now represented in the sequenced clones and more than 94% is represented in the combined publicly available sequence databases. The total amount of finished sequence is now around 1 Gb.

Broad genomic landscape

What biological insights can be gleaned from the draft sequence? In this section, we consider very large-scale features of the draft genome sequence: the distribution of GC content, CpG islands and recombination rates, and the repeat content and gene content of the human genome. The draft genome sequence makes it possible to integrate these features and others at scales ranging from individual

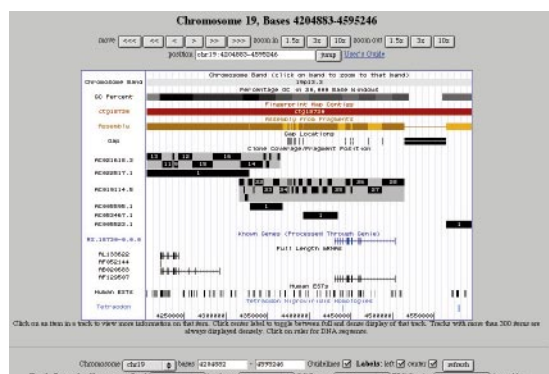


Figure 10 Screen shot from the UCSC Draft Human Genome Browser. See <http://genome.ucsc.edu/>.

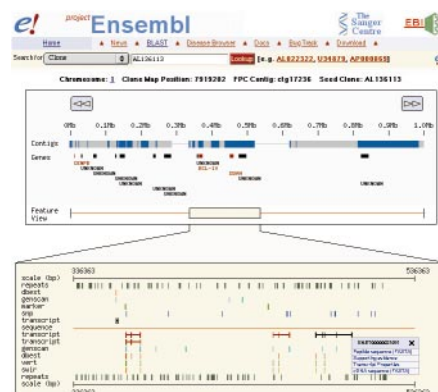


Figure 11 Screen shot from the Genome Browser of Project Ensembl. See <http://www.ensembl.org>.

nucleotides to collections of chromosomes. Unless noted, all analyses were conducted on the assembled draft genome sequence described above.

Figure 9 provides a high-level view of the contents of the draft genome sequence, at a scale of about 3.8 Mb per centimetre. Of course, navigating information spanning nearly ten orders of magnitude requires computational tools to extract the full value. We have created and made freely available various 'Genome Browsers'. Browsers were developed and are maintained by the University of California at Santa Cruz (Fig. 10) and the EnSEMBL project of the European Bioinformatics Institute and the Sanger Centre (Fig. 11). Additional browsers have been created; URLs are listed at www.nhgri.nih.gov/genome_hub. These web-based computer tools allow users to view an annotated display of the draft genome sequence, with the ability to scroll along the chromosomes and zoom in or out to different scales. They include: the nucleotide sequence, sequence contigs, clone contigs, sequence coverage and finishing status, local GC content, CpG islands, known STS markers from previous genetic and physical maps, families of repeat sequences, known genes, ESTs and mRNAs, predicted genes, SNPs and sequence similarities with other organisms (currently the pufferfish *Tetraodon nigroviridis*). These browsers will be updated as the draft genome sequence is refined and corrected as additional annotations are developed.

In addition to using the Genome Browsers, one can download

Box 2

Sources of publicly available sequence data and other relevant genomic information

http://genome.ucsc.edu/ University of California at Santa Cruz Contains the assembly of the draft genome sequence used in this paper and updates
http://genome.wustl.edu/gsc/human/Mapping/ Washington University Contains links to clone and accession maps of the human genome
http://www.ensembl.org EBI/Sanger Centre Allows access to DNA and protein sequences with automatic baseline annotation
http://www.ncbi.nlm.nih.gov/genome/guide/ NCBI Views of chromosomes and maps and loci with links to other NCBI resources
http://www.ncbi.nlm.nih.gov/genemap99/ Gene map 99: contains data and viewers for radiation hybrid maps of EST-based STSs
http://compbio.ornl.gov/channel/index.html Oak Ridge National Laboratory Java viewers for human genome data
http://hgrep.ims.u-tokyo.ac.jp/ RIKEN and the University of Tokyo Gives an overview of the entire human genome structure
http://snp.cshl.org/ The SNP Consortium Includes a variety of ways to query for SNPs in the human genome
http://www.ncbi.nlm.nih.gov/Omim/ Online Mendelian Inheritance in Man Contain information about human genes and disease
http://www.nhgri.nih.gov/ELSI/ and http://www.ornl.gov/hgmis/elsi/elsi.html NHGRI and DOE Contains information, links and articles on a wide range of social, ethical and legal issues

from these sites the entire draft genome sequence together with the annotations in a computer-readable format. The sequences of the underlying sequenced clones are all available through the public sequence databases. URLs for these and other genome websites are listed in Box 2. A larger list of useful URLs can be found at www.nhgri.nih.gov/genome_hub. An introduction to using the draft genome sequence, as well as associated databases and analytical tools, is provided in an accompanying paper¹¹¹.

In addition, the human cytogenetic map has been integrated with the draft genome sequence as part of a related project. The BAC Resource Consortium¹⁰³ established dense connections between the maps using more than 7,500 sequenced large-insert clones that had been cytogenetically mapped by FISH; the average density of the map is 2.3 clones per Mb. Although the precision of the integration is limited by the resolution of FISH, the links provide a powerful tool for the analysis of cytogenetic aberrations in inherited diseases and cancer. These cytogenetic links can also be accessed through the Genome Browsers.

Long-range variation in GC content

The existence of GC-rich and GC-poor regions in the human genome was first revealed by experimental studies involving density gradient separation, which indicated substantial variation in average GC content among large fragments. Subsequent studies have indicated that these GC-rich and GC-poor regions may have different biological properties, such as gene density, composition of repeat sequences, correspondence with cytogenetic bands and recombination rate^{112–117}. Many of these studies were indirect, owing to the lack of sufficient sequence data.

The draft genome sequence makes it possible to explore the variation in GC content in a direct and global manner. Visual inspection (Fig. 9) confirms that local GC content undergoes substantial long-range excursions from its genome-wide average of 41%. If the genome were drawn from a uniform distribution of GC content, the local GC content in a window of size n bp should be $41 \pm \sqrt{((41)(59)/n)\%}$. Fluctuations would be modest, with the standard deviation being halved as the window size is quadrupled—for example, 0.70%, 0.35%, 0.17% and 0.09% for windows of size 5, 20, 80 and 320 kb.

The draft genome sequence, however, contains many regions with much more extreme variation. There are huge regions (> 10 Mb) with GC content far from the average. For example, the most distal 48 Mb of chromosome 1p (from the telomere to about STS marker D1S3279) has an average GC content of 47.1%, and chromosome 13 has a 40-Mb region (roughly between STS marker A005X38 and

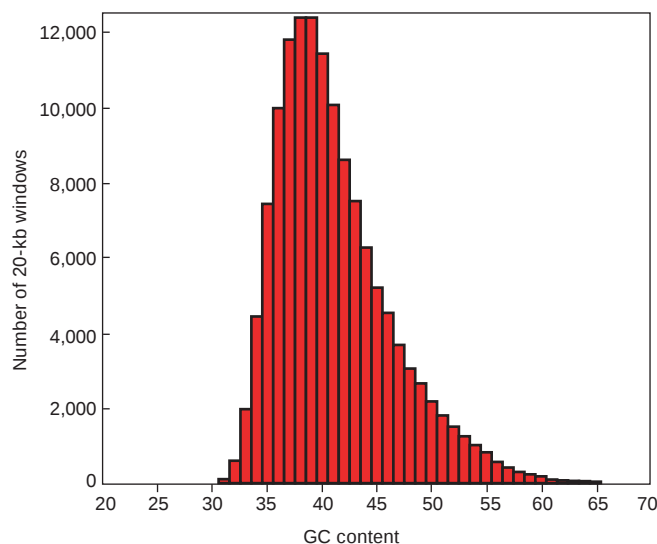


Figure 12 Histogram of GC content of 20-kb windows in the draft genome sequence.

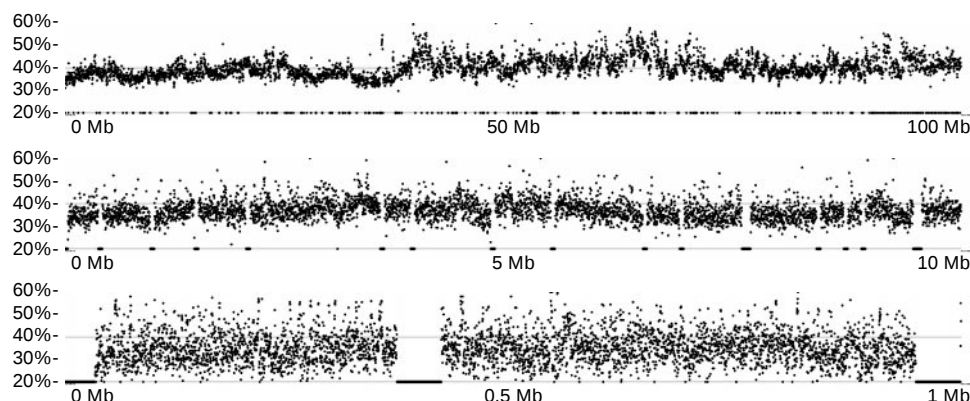


Figure 13 Variation in GC content at various scales. The GC content in subregions of a 100-Mb region of chromosome 1 is plotted, starting at about 83 Mb from the beginning of the draft genome sequence. This region is AT-rich overall. Top, the GC content of the

entire 100-Mb region analysed in non-overlapping 20-kb windows. Middle, GC content of the first 10 Mb, analysed in 2-kb windows. Bottom, GC content of the first 1 Mb, analysed in 200-bp windows. At this scale, gaps in the sequence can be seen.

stsG30423) with only 36% GC content. There are also examples of large shifts in GC content between adjacent multimegabase regions. For example, the average GC content on chromosome 17q is 50% for the distal 10.3 Mb but drops to 38% for the adjacent 3.9 Mb. There are regions of less than 300 kb with even wider swings in GC content, for example, from 33.1% to 59.3%.

Long-range variation in GC content is evident not just from extreme outliers, but throughout the genome. The distribution of average GC content in 20-kb windows across the draft genome sequence is shown in Fig. 12. The spread is 15-fold larger than predicted by a uniform process. Moreover, the standard deviation barely decreases as window size increases by successive factors of four—5.9%, 5.2%, 4.9% and 4.6% for windows of size 5, 20, 80 and 320 kb. The distribution is also notably skewed, with 58% below the average and 42% above the average of 41%, with a long tail of GC-rich regions.

Bernardi and colleagues^{118,119} proposed that the long-range variation in GC content may reflect that the genome is composed of a mosaic of compositionally homogeneous regions that they dubbed 'isochores'. They suggested that the skewed distribution is composed of five normal distributions, corresponding to five distinct types of isochore (L1, L2, H1, H2 and H3, with GC contents of < 38%, 38–42%, 42–47%, 47–52% and > 52%, respectively).

We studied the draft genome sequence to see whether strict isochores could be identified. For example, the sequence was divided into 300-kb windows, and each window was subdivided into 20-kb subwindows. We calculated the average GC content for each window and subwindow, and investigated how much of the variance in the GC content of subwindows across the genome can be statistically 'explained' by the average GC content in each window. About three-quarters of the genome-wide variance among 20-kb windows can be statistically explained by the average GC content of 300-kb windows that contain them, but the residual variance among subwindows (standard deviation, 2.4%) is still far too large to be consistent with a homogeneous distribution. In fact, the hypothesis of homogeneity could be rejected for each 300-kb window in the draft genome sequence.

Similar results were obtained with other window and subwindow sizes. Some of the local heterogeneity in GC content is attributable to transposable element insertions (see below). Such repeat elements typically have a higher GC content than the surrounding sequence, with the effect being strongest for the most recent insertions.

These results rule out a strict notion of isochores as compositionally homogeneous. Instead, there is substantial variation at many different scales, as illustrated in Fig. 13. Although isochores do not appear to merit the prefix 'iso', the genome clearly does contain large regions of distinctive GC content and it is likely to be

worth redefining the concept so that it becomes possible rigorously to partition the genome into regions. In the absence of a precise definition, we will loosely refer to such regions as 'GC content domains' in the context of the discussion below.

Fickett *et al.*¹²⁰ have explored a model in which the underlying preference for a particular GC content drifts continuously throughout the genome, an approach that bears further examination. Churchill¹²¹ has proposed that the boundaries between GC content domains can in some cases be predicted by a hidden Markov model, with one state representing a GC-rich region and one representing an AT-rich region. We found that this approach tended to identify only very short domains of less than a kilobase (data not shown), but variants of this approach deserve further attention.

The correlation between GC content domains and various biological properties is of great interest, and this is likely to be the most fruitful route to understanding the basis of variation in GC content. As described below, we confirm the existence of strong correlations with both repeat content and gene density. Using the integration between the draft genome sequence and the cytogenetic map described above, it is possible to confirm a statistically significant correlation between GC content and Giemsa bands (G-bands). For example, 98% of large-insert clones mapping to the darkest G-bands are in 200-kb regions of low GC content (average 37%), whereas more than 80% of clones mapping to the lightest G-bands are in regions of high GC content (average 45%)¹⁰³. Estimated band locations can be seen in Fig. 9 and viewed in the context of other genome annotation at <http://genome.ucsc.edu/goldenPath/mapPlots/> and <http://genome.ucsc.edu/goldenPath/hgTracks.html>.

CpG islands

A related topic is the distribution of so-called CpG islands across the genome. The dinucleotide CpG is notable because it is greatly under-represented in human DNA, occurring at only about one-fifth of the roughly 4% frequency that would be expected by simply multiplying the typical fraction of Cs and Gs (0.21×0.21). The deficit occurs because most CpG dinucleotides are methylated on the cytosine base, and spontaneous deamination of methyl-C residues gives rise to T residues. (Spontaneous deamination of ordinary cytosine residues gives rise to uracil residues that are readily recognized and repaired by the cell.) As a result, methyl-CpG dinucleotides steadily mutate to TpG dinucleotides. However, the genome contains many 'CpG islands' in which CpG dinucleotides are not methylated and occur at a frequency closer to that predicted by the local GC content. CpG islands are of particular interest because many are associated with the 5' ends of genes^{122–127}.

We searched the draft genome sequence for CpG islands. Ideally, they should be defined by directly testing for the absence of cytosine methylation, but that was not practical for this report. There are

Table 10 Number of CpG islands by GC content

GC content of island	Number of islands	Percentage of islands	Nucleotides in islands	Percentage of nucleotides in islands
Total	28,890	100	19,818,547	100
>80%	22	0.08	5,916	0.03
70–80%	5,884	20	3,111,965	16
60–70%	18,779	65	13,110,924	66
50–60%	4,205	15	3,589,742	18

Potential CpG islands were identified by searching the draft genome sequence one base at a time, scoring each dinucleotide (+17 for GC, –1 for others) and identifying maximally scoring segments. Each segment was then evaluated to determine GC content ($\geq 50\%$), length (>200) and ratio of observed proportion of GC dinucleotides to the expected proportion on the basis of the GC content of the segment (>0.60), using a modification of a program developed by G. Micklem (personal communication).

various computer programs that attempt to identify CpG islands on the basis of primary sequence alone. These programs differ in some important respects (such as how aggressively they subdivide long CpG-containing regions), and the precise correspondence with experimentally undermethylated islands has not been validated. Nevertheless, there is a good correlation, and computational analysis thus provides a reasonable picture of the distribution of CpG islands in the genome.

To identify CpG islands, we used the definition proposed by Gardiner-Garden and Frommer¹²⁸ and embodied in a computer program. We searched the draft genome sequence for CpG islands, using both the full sequence and the sequence masked to eliminate repeat sequences. The number of regions satisfying the definition of a CpG island was 50,267 in the full sequence and 28,890 in the repeat-masked sequence. The difference reflects the fact that some repeat elements (notably Alu) are GC-rich. Although some of these repeat elements may function as control regions, it seems unlikely that most of the apparent CpG islands in repeat sequences are functional. Accordingly, we focused on those in the non-repeated sequence. The count of 28,890 CpG islands is reasonably close to the previous estimate of about 35,000 (ref. 129, as modified by ref. 130). Most of the islands are short, with 60–70% GC content (Table 10). More than 95% of the islands are less than 1,800 bp long, and more than 75% are less than 850 bp. The longest CpG island (on chromosome 10) is 36,619 bp long, and 322 are longer than 3,000 bp. Some of the larger islands contain ribosomal pseudogenes, although RNA genes and pseudogenes account for only a small proportion of all islands ($<0.5\%$). The smaller islands are consistent with their previously hypothesized function, but the role of these larger islands is uncertain.

The density of CpG islands varies substantially among some of the chromosomes. Most chromosomes have 5–15 islands per Mb, with a mean of 10.5 islands per Mb. However, chromosome Y has an

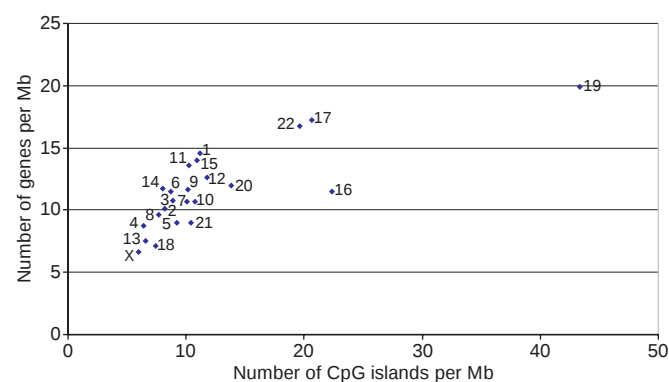


Figure 14 Number of CpG islands per Mb for each chromosome, plotted against the number of genes per Mb (the number of genes was taken from GeneMap98 (ref. 100)). Chromosomes 16, 17, 22 and particularly 19 are clear outliers, with a density of CpG islands that is even greater than would be expected from the high gene counts for these four chromosomes.

unusually low 2.9 islands per Mb, and chromosomes 16, 17 and 22 have 19–22 islands per Mb. The extreme outlier is chromosome 19, with 43 islands per Mb. Similar trends are seen when considering the percentage of bases contained in CpG islands. The relative density of CpG islands correlates reasonably well with estimates of relative gene density on these chromosomes, based both on previous mapping studies involving ESTs (Fig. 14) and on the distribution of gene predictions discussed below.

Comparison of genetic and physical distance

The draft genome sequence makes it possible to compare genetic and physical distances and thereby to explore variation in the rate of recombination across the human chromosomes. We focus here on large-scale variation. Finer variation is examined in an accompanying paper¹³¹.

The genetic and physical maps are integrated by 5,282 polymorphic loci from the Marshfield genetic map¹⁰², whose positions are known in terms of centimorgans (cM) and Mb along the chromosomes. Figure 15 shows the comparison of the draft genome sequence for chromosome 12 with the male, female and sex-averaged maps. One can calculate the approximate ratio of cM per Mb across a chromosome (reflected in the slopes in Fig. 15) and the average recombination rate for each chromosome arm.

Two striking features emerge from analysis of these data. First, the average recombination rate increases as the length of the chromosome arm decreases (Fig. 16). Long chromosome arms have an average recombination rate of about 1 cM per Mb, whereas the shortest arms are in the range of 2 cM per Mb. A similar trend has been seen in the yeast genome^{132,133}, despite the fact that the physical scale is nearly 200 times as small. Moreover, experimental studies have shown that lengthening or shortening yeast chromosomes results in a compensatory change in recombination rate¹³².

The second observation is that the recombination rate tends to be suppressed near the centromeres and higher in the distal portions of most chromosomes, with the increase largely in the terminal

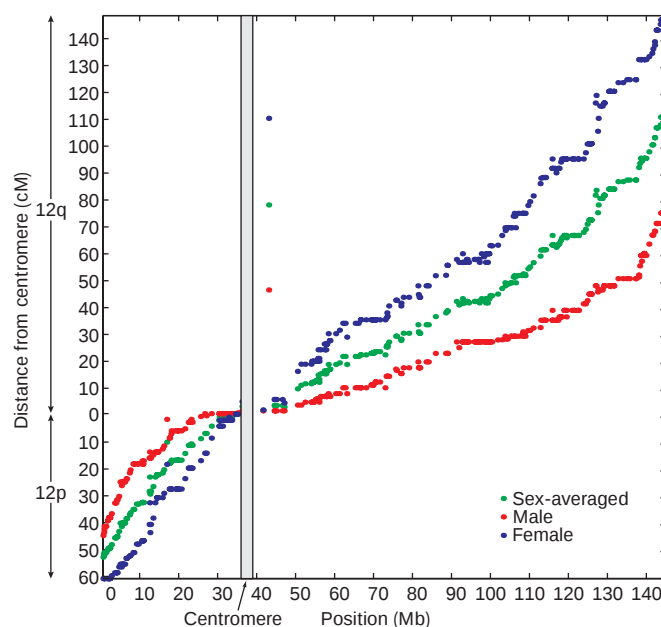


Figure 15 Distance in cM along the genetic map of chromosome 12 plotted against position in Mb in the draft genome sequence. Female, male and sex-averaged maps are shown. Female recombination rates are much higher than male recombination rates. The increased slopes at either end of the chromosome reflect the increased rates of recombination per Mb near the telomeres. Conversely, the flatter slope near the centromere shows decreased recombination there, especially in male meiosis. This is typical of the other chromosomes as well (see <http://genome.ucsc.edu/goldenPath/mapPlots>). Discordant markers may be map, marker placement or assembly errors.

20–35 Mb. The increase is most pronounced in the male meiotic map. The effect can be seen, for example, from the higher slope at both ends of chromosome 12 (Fig. 15). Regional and sex-specific effects have been observed for chromosome 21 (refs 110, 134).

Why is recombination higher on smaller chromosome arms? A higher rate would increase the likelihood of at least one crossover during meiosis on each chromosome arm, as is generally observed in human chiasmata counts¹³⁵. Crossovers are believed to be necessary for normal meiotic disjunction of homologous chromosome pairs in eukaryotes. An extreme example is the pseudoautosomal regions on chromosomes Xp and Yp, which pair during male meiosis; this physical region of only 2.6 Mb has a genetic length of 50 cM (corresponding to 20 cM per Mb), with the result that a crossover is virtually assured.

Mechanistically, the increased rate of recombination on shorter chromosome arms could be explained if, once an initial recombination event occurs, additional nearby events are blocked by positive crossover interference on each arm. Evidence from yeast mutants in which interference is abolished shows that interference plays a key role in distributing a limited number of crossovers among the various chromosome arms in yeast¹³⁶. An alternative possibility is that a checkpoint mechanism scans for and enforces the presence of at least one crossover on each chromosome arm.

Variation in recombination rates along chromosomes and between the sexes is likely to reflect variation in the initiation of meiosis-induced double-strand breaks (DSBs) that initiate recombination. DSBs in yeast have been associated with open chromatin^{137,138}, rather than with specific DNA sequence motifs. With the availability of the draft genome sequence, it should be possible to explore in an analogous manner whether variation in human recombination rates reflects systematic differences in chromosome accessibility during meiosis.

Repeat content of the human genome

A puzzling observation in the early days of molecular biology was that genome size does not correlate well with organismal complexity. For example, *Homo sapiens* has a genome that is 200 times as large as that of the yeast *S. cerevisiae*, but 200 times as small as that of

Amoeba dubia^{139,140}. This mystery (the C-value paradox) was largely resolved with the recognition that genomes can contain a large quantity of repetitive sequence, far in excess of that devoted to protein-coding genes (reviewed in refs 140, 141).

In the human, coding sequences comprise less than 5% of the genome (see below), whereas repeat sequences account for at least 50% and probably much more. Broadly, the repeats fall into five classes: (1) transposon-derived repeats, often referred to as interspersed repeats; (2) inactive (partially) retroposed copies of cellular genes (including protein-coding genes and small structural RNAs), usually referred to as processed pseudogenes; (3) simple sequence repeats, consisting of direct repetitions of relatively short *k*-mers such as (A)_{*n*}, (CA)_{*n*} or (CGG)_{*n*}; (4) segmental duplications, consisting of blocks of around 10–300 kb that have been copied from one region of the genome into another region; and (5) blocks of tandemly repeated sequences, such as at centromeres, telomeres, the short arms of acrocentric chromosomes and ribosomal gene clusters. (These regions are intentionally under-represented in the draft genome sequence and are not discussed here.)

Repeats are often described as ‘junk’ and dismissed as uninteresting. However, they actually represent an extraordinary trove of information about biological processes. The repeats constitute a rich palaeontological record, holding crucial clues about evolutionary events and forces. As passive markers, they provide assays for studying processes of mutation and selection. It is possible to recognize cohorts of repeats ‘born’ at the same time and to follow their fates in different regions of the genome or in different species. As active agents, repeats have reshaped the genome by causing ectopic rearrangements, creating entirely new genes, modifying and reshuffling existing genes, and modulating overall GC content. They also shed light on chromosome structure and dynamics, and provide tools for medical genetic and population genetic studies.

The human is the first repeat-rich genome to be sequenced, and so we investigated what information could be gleaned from this majority component of the human genome. Although some of the general observations about repeats were suggested by previous studies, the draft genome sequence provides the first comprehensive view, allowing some questions to be resolved and new mysteries to emerge.

Transposon-derived repeats

Most human repeat sequence is derived from transposable elements^{142,143}. We can currently recognize about 45% of the genome as belonging to this class. Much of the remaining ‘unique’ DNA must also be derived from ancient transposable element copies that have diverged too far to be recognized as such. To describe our analyses of interspersed repeats, it is necessary briefly to review the relevant features of human transposable elements.

Classes of transposable elements. In mammals, almost all transposable elements fall into one of four types (Fig. 17), of which three transpose through RNA intermediates and one transposes directly as DNA. These are long interspersed elements (LINEs), short interspersed elements (SINEs), LTR retrotransposons and DNA transposons.

LINEs are one of the most ancient and successful inventions in eukaryotic genomes. In humans, these transposons are about 6 kb long, harbour an internal polymerase II promoter and encode two open reading frames (ORFs). Upon translation, a LINE RNA assembles with its own encoded proteins and moves to the nucleus, where an endonuclease activity makes a single-stranded nick and the reverse transcriptase uses the nicked DNA to prime reverse transcription from the 3′ end of the LINE RNA. Reverse transcription frequently fails to proceed to the 5′ end, resulting in many truncated, nonfunctional insertions. Indeed, most LINE-derived repeats are short, with an average size of 900 bp for all LINE1 copies, and a median size of 1,070 bp for copies of the currently active LINE1 element (L1Hs). New insertion sites are flanked by a small

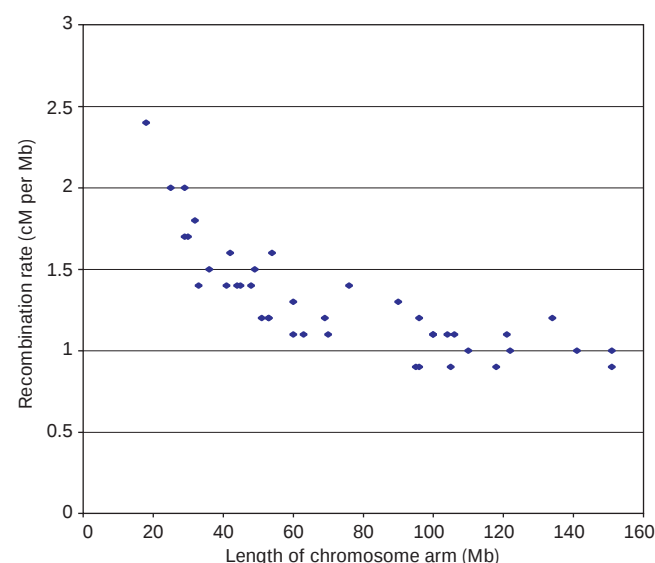


Figure 16 Rate of recombination averaged across the euchromatic portion of each chromosome arm plotted against the length of the chromosome arm in Mb. For large chromosomes, the average recombination rates are very similar, but as chromosome arm length decreases, average recombination rates rise markedly.

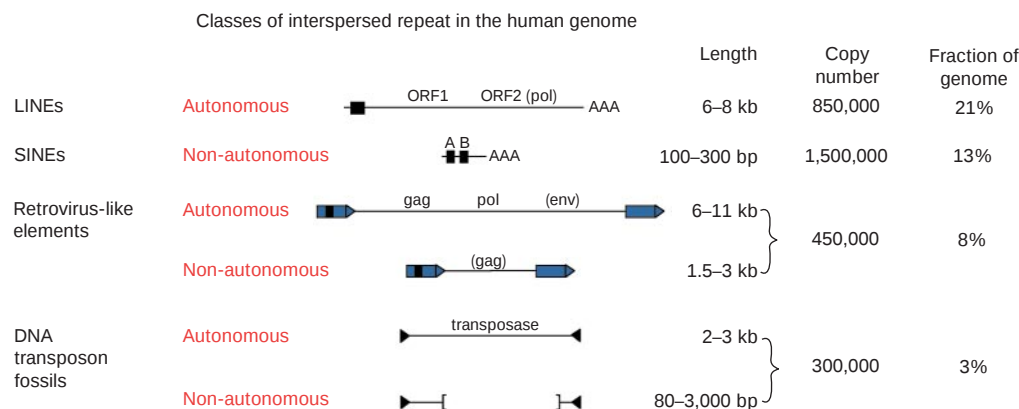


Figure 17 Almost all transposable elements in mammals fall into one of four classes. See text for details.

target site duplication of 7–20 bp. The LINE machinery is believed to be responsible for most reverse transcription in the genome, including the retrotransposition of the non-autonomous SINEs¹⁴⁴ and the creation of processed pseudogenes^{145,146}. Three distantly related LINE families are found in the human genome: LINE1, LINE2 and LINE3. Only LINE1 is still active.

SINEs are wildly successful freeloaders on the backs of LINE elements. They are short (about 100–400 bp), harbour an internal polymerase III promoter and encode no proteins. These non-autonomous transposons are thought to use the LINE machinery for transposition. Indeed, most SINEs ‘live’ by sharing the 3’ end with a resident LINE element¹⁴⁴. The promoter regions of all known SINEs are derived from tRNA sequences, with the exception of a single monophyletic family of SINEs derived from the signal recognition particle component 7SL. This family, which also does not share its 3’ end with a LINE, includes the only active SINE in the human genome: the Alu element. By contrast, the mouse has both tRNA-derived and 7SL-derived SINEs. The human genome contains three distinct monophyletic families of SINEs: the active Alu, and the inactive MIR and Ther2/MIR3.

LTR retroposons are flanked by long terminal direct repeats that contain all of the necessary transcriptional regulatory elements. The autonomous elements (retrotransposons) contain *gag* and *pol* genes, which encode a protease, reverse transcriptase, RNase H and integrase. Exogenous retroviruses seem to have arisen from endogenous retrotransposons by acquisition of a cellular *envelope* gene (*env*)¹⁴⁷. Transposition occurs through the retroviral mechanism with reverse transcription occurring in a cytoplasmic virus-like particle, primed by a tRNA (in contrast to the nuclear location and chromosomal priming of LINEs). Although a variety of LTR retrotransposons exist, only the vertebrate-specific endogenous retroviruses (ERVs) appear to have been active in the mammalian genome. Mammalian retroviruses fall into three classes (I–III), each comprising many families with independent origins. Most (85%) of the LTR retroposon-derived ‘fossils’ consist only of an isolated LTR, with the internal sequence having been lost by homologous recombination between the flanking LTRs.

DNA transposons resemble bacterial transposons, having terminal inverted repeats and encoding a transposase that binds near the inverted repeats and mediates mobility through a ‘cut-and-paste’ mechanism. The human genome contains at least seven major classes of DNA transposon, which can be subdivided into many families with independent origins¹⁴⁸ (see RepBase, <http://www.girinst.org/~server/repbase.html>). DNA transposons tend to have short life spans within a species. This can be explained by contrasting the modes of transposition of DNA transposons and LINE elements. LINE transposition tends to involve only functional elements, owing to the *cis*-preference by which LINE proteins assemble with the RNA from which they were translated. By

contrast, DNA transposons cannot exercise a *cis*-preference: the encoded transposase is produced in the cytoplasm and, when it returns to the nucleus, it cannot distinguish active from inactive elements. As inactive copies accumulate in the genome, transposition becomes less efficient. This checks the expansion of any DNA transposon family and in due course causes it to die out. To survive, DNA transposons must eventually move by horizontal transfer to virgin genomes, and there is considerable evidence for such transfer^{149–153}.

Transposable elements employ different strategies to ensure their evolutionary survival. LINEs and SINEs rely almost exclusively on vertical transmission within the host genome¹⁵⁴ (but see refs 148, 155). DNA transposons are more promiscuous, requiring relatively frequent horizontal transfer. LTR retroposons use both strategies, with some being long-term active residents of the human genome (such as members of the ERVL family) and others having only short residence times.

Table 11 Number of copies and fraction of genome for classes of interspersed repeat

	Number of copies (x 1,000)	Total number of bases in the draft genome sequence (Mb)	Fraction of the draft genome sequence (%)	Number of families (subfamilies)
SINEs	1,558	359.6	13.14	
Alu	1,090	290.1	10.60	1 (~20)
MIR	393	60.1	2.20	1 (1)
MIR3	75	9.3	0.34	1 (1)
LINEs	868	558.8	20.42	
LINE1	516	462.1	16.89	1 (~55)
LINE2	315	88.2	3.22	1 (2)
LINE3	37	8.4	0.31	1 (2)
LTR elements	443	227.0	8.29	
ERV-class I	112	79.2	2.89	72 (132)
ERV(K)-class II	8	8.5	0.31	10 (20)
ERV (L)-class III	83	39.5	1.44	21 (42)
MaLR	240	99.8	3.65	1 (31)
DNA elements	294	77.6	2.84	
hAT group				
MER1-Charlie	182	38.1	1.39	25 (50)
Zaphod	13	4.3	0.16	4 (10)
Tc-1 group				
MER2-Tigger	57	28.0	1.02	12 (28)
Tc2	4	0.9	0.03	1 (5)
Mariner	14	2.6	0.10	4 (5)
PiggyBac-like	2	0.5	0.02	10 (20)
Unclassified	22	3.2	0.12	7 (7)
Unclassified	3	3.8	0.14	3 (4)
Total interspersed repeats		1,226.8	44.83	

The number of copies and base pair contributions of the major classes and subclasses of transposable elements in the human genome. Data extracted from a RepeatMasker analysis of the draft genome sequence (RepeatMasker version 09092000, sensitive settings, using RepBase Update 5.08). In calculating percentages, RepeatMasker excluded the runs of Ns linking the contigs in the draft genome sequence. In the last column, separate consensus sequences in the repeat databases are considered subfamilies, rather than families, when the sequences are closely related or related through intermediate subfamilies.

Census of human repeats. We began by taking a census of the transposable elements in the draft genome sequence, using a recently updated version of the RepeatMasker program (version 09092000) run under sensitive settings (see <http://repeatmasker.genome.washington.edu>). This program scans sequences to identify full-length and partial members of all known repeat families represented in RepBase Update (version 5.08; see <http://www.girinst.org/~server/repbase.html> and ref. 156). Table 11 shows the number of copies and fraction of the draft genome sequence occupied by each of the four major classes and the main subclasses.

The precise count of repeats is obviously underestimated because the genome sequence is not finished, but their density and other properties can be stated with reasonable confidence. Currently recognized SINEs, LINEs, LTR retrotransposons and DNA transposon copies comprise 13%, 20%, 8% and 3% of the sequence, respectively. We expect these densities to grow as more repeat families are recognized, among which will be lower copy number LTR elements and DNA transposons, and possibly high copy number ancient (highly diverged) repeats.

Age distribution. The age distribution of the repeats in the human genome provides a rich 'fossil record' stretching over several hundred million years. The ancestry and approximate age of each fossil can be inferred by exploiting the fact that each copy is derived from, and therefore initially carried the sequence of, a then-active

transposon and, being generally under no functional constraint, has accumulated mutations randomly and independently of other copies. We can infer the sequence of the ancestral active elements by clustering the modern derivatives into phylogenetic trees and building a consensus based on the multiple sequence alignment of a cluster of copies. Using available consensus sequences for known repeat subfamilies, we calculated the per cent divergence from the inferred ancestral active transposon for each of three million interspersed repeats in the draft genome sequence.

The percentage of sequence divergence can be converted into an approximate age in millions of years (Myr) on the basis of evolutionary information. Care is required in calibrating the clock, because the rate of sequence divergence may not be constant over time or between lineages¹³⁹. The relative-rate test¹⁵⁷ can be used to calculate the sequence divergence that accumulated in a lineage after a given timepoint, on the basis of comparison with a sibling species that diverged at that time and an outgroup species. For example, the substitution rate over roughly the last 25 Myr in the human lineage can be calculated by using old world monkeys (which diverged about 25 Myr ago) as a sibling species and new world monkeys as an outgroup. We have used currently available calibrations for the human lineage, but the issue should be revisited as sequence information becomes available from different mammals.

Figure 18a shows the representation of various classes of trans-

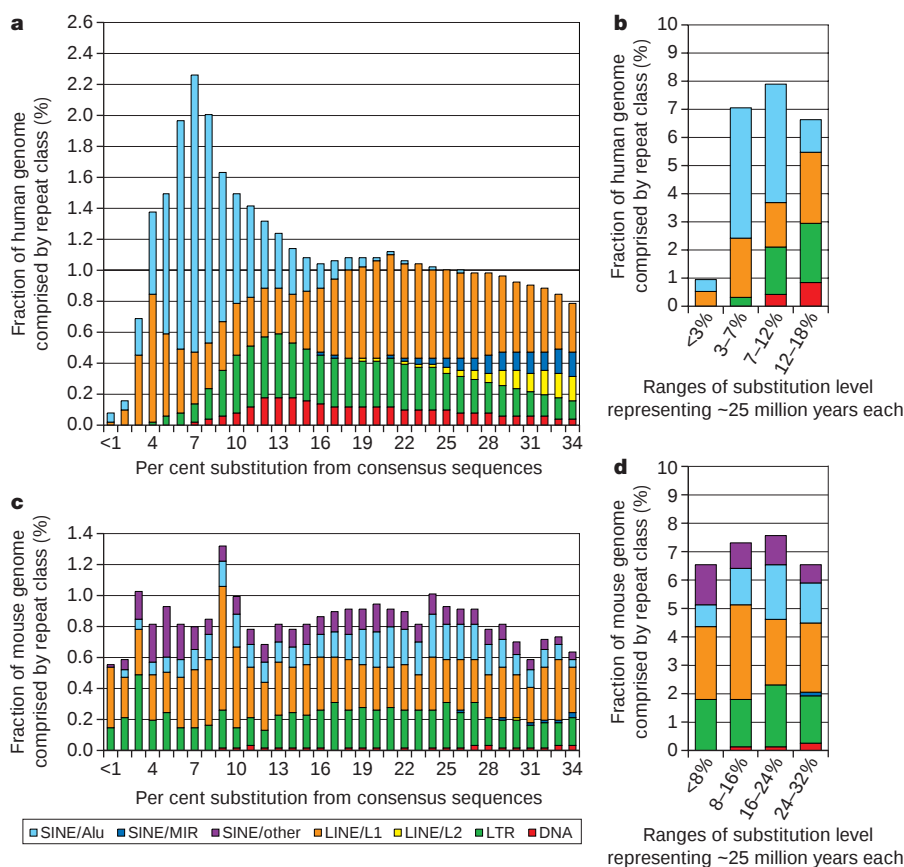


Figure 18 Age distribution of interspersed repeats in the human and mouse genomes. Bases covered by interspersed repeats were sorted by their divergence from their consensus sequence (which approximates the repeat's original sequence at the time of insertion). The average number of substitutions per 100 bp (substitution level, K) was calculated from the mismatch level p assuming equal frequency of all substitutions (the one-parameter Jukes-Cantor model, $K = -3/4 \ln(1 - 4/3p)$). This model tends to underestimate higher substitution levels. CpG dinucleotides in the consensus were excluded from the substitution level calculations because the C→T transition rate in CpG pairs is about tenfold higher than other transitions and causes distortions in comparing

transposable elements with high and low CpG content. **a**, The distribution, for the human genome, in bins corresponding to 1% increments in substitution levels. **b**, The data grouped into bins representing roughly equal time periods of 25 Myr. **c,d**, Equivalent data for available mouse genomic sequence. There is a different correspondence between substitution levels and time periods owing to different rates of nucleotide substitution in the two species. The correspondence between substitution levels and time periods was largely derived from three-way species comparisons (relative rate test^{139,157}) with the age estimates based on fossil data. Human divergence from gibbon 20–30 Myr; old world monkey 25–35 Myr; prosimians 55–80 Myr; eutherian mammalian radiation ~100 Myr.

possible elements in categories reflecting equal amounts of sequence divergence. In Fig. 18b the data are grouped into four bins corresponding to successive 25-Myr periods, on the basis of an approximate clock. Figure 19 shows the mean ages of various subfamilies of DNA transposons. Several facts are apparent from these graphs. First, most interspersed repeats in the human genome predate the eutherian radiation. This is a testament to the extremely slow rate with which nonfunctional sequences are cleared from vertebrate genomes (see below concerning comparison with the fly).

Second, LINE and SINE elements have extremely long lives. The monophyletic LINE1 and Alu lineages are at least 150 and 80 Myr old, respectively. In earlier times, the reigning transposons were LINE2 and MIR^{148,158}. The SINE MIR was perfectly adapted for reverse transcription by LINE2, as it carried the same 50-base sequence at its 3' end. When LINE2 became extinct 80–100 Myr ago, it spelled the doom of MIR.

Third, there were two major peaks of DNA transposon activity (Fig. 19). The first involved Charlie elements and occurred long before the eutherian radiation; the second involved Tigger elements and occurred after this radiation. Because DNA transposons can produce large-scale chromosome rearrangements^{159–162}, it is possible that widespread activity could be involved in speciation events.

Fourth, there is no evidence for DNA transposon activity in the past 50 Myr in the human genome. The youngest two DNA transposon families that we can identify in the draft genome sequence (MER75 and MER85) show 6–7% divergence from their respective consensus sequences representing the ancestral element (Fig. 19), indicating that they were active before the divergence of humans and new world monkeys. Moreover, these elements were relatively unsuccessful, together contributing just 125 kb to the draft genome sequence.

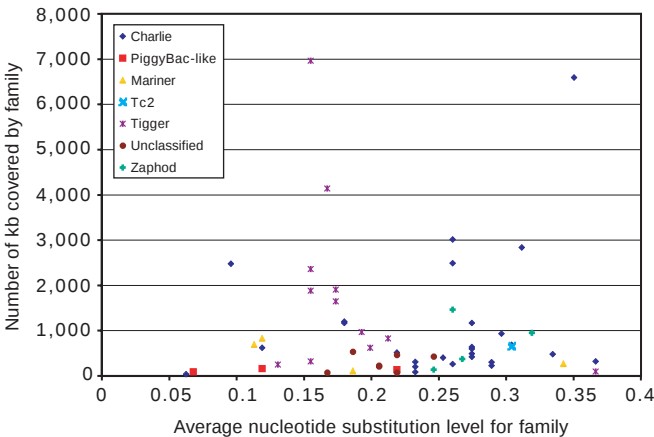


Figure 19 Median ages and per cent of the genome covered by subfamilies of DNA transposons. The Charlie and Zaphod elements were hobo-Activator-Tam3 (hAT) DNA transposons; Mariner, Tc2 and Tigger were Tc1-like elements. Unlike retroposons, DNA transposons are thought to have a short life span in a genome. Thus, the average or median divergence of copies from the consensus is a particularly accurate measure of the age of the DNA transposon copies.

Finally, LTR retroposons appear to be teetering on the brink of extinction, if they have not already succumbed. For example, the most prolific elements (ERVL and MaLRs) flourished for more than 100 Myr but appear to have died out about 40 Myr ago^{163,164}. Only a single LTR retroposon family (HERVK10) is known to have transposed since our divergence from the chimpanzee 7 Myr ago, with only one known copy (in the HLA region) that is not shared between all humans¹⁶⁵. In the draft genome sequence, we can identify only three full-length copies with all ORFs intact (the final total may be slightly higher owing to the imperfect state of the draft genome sequence).

More generally, the overall activity of all transposons has declined markedly over the past 35–50 Myr, with the possible exception of LINE1 (Fig. 18). Indeed, apart from an exceptional burst of activity of Alus peaking around 40 Myr ago, there would appear to have been a fairly steady decline in activity in the hominid lineage since the mammalian radiation. The extent of the decline must be even greater than it appears because old repeats are gradually removed by random deletion and because old repeat families are harder to recognize and likely to be under-represented in the repeat databases. (We confirmed that the decline in transposition is not an artefact arising from errors in the draft genome sequence, which, in principle, could increase the divergence level in recent elements. First, the sequence error rate (Table 9) is far too low to have a significant effect on the apparent age of recent transposons; and second, the same result is seen if one considers only finished sequence.)

What explains the decline in transposon activity in the lineage leading to humans? We return to this question below, in the context of the observation that there is no similar decline in the mouse genome.

Comparison with other organisms. We compared the complement of transposable elements in the human genome with those of the other sequenced eukaryotic genomes. We analysed the fly, worm and mustard weed genomes for the number and nature of repeats (Table 12) and the age distribution (Fig. 20). (For the fly, we analysed the 114 Mb of unfinished ‘large’ contigs produced by the whole-genome shotgun assembly¹⁶⁶, which are reported to represent euchromatic sequence. Similar results were obtained by analysing 30 Mb of finished euchromatic sequence.) The human genome stands in stark contrast to the genomes of the other organisms.

(1) The euchromatic portion of the human genome has a much higher density of transposable element copies than the euchromatic DNA of the other three organisms. The repeats in the other organisms may have been slightly underestimated because the repeat databases for the other organisms are less complete than for the human, especially with regard to older elements; on the other hand, recent additions to these databases appear to increase the repeat content only marginally.

(2) The human genome is filled with copies of ancient transposons, whereas the transposons in the other genomes tend to be of more recent origin. The difference is most marked with the fly, but is clear for the other genomes as well. The accumulation of old repeats is likely to be determined by the rate at which organisms engage in ‘housecleaning’ through genomic deletion. Studies of pseudogenes

Table 12 Number and nature of interspersed repeats in eukaryotic genomes

	Human		Fly		Worm		Mustard weed	
	Percentage of bases	Approximate number of families	Percentage of bases	Approximate number of families	Percentage of bases	Approximate number of families	Percentage of bases	Approximate number of families
LINE/SINE	33.40%	6	0.70%	20	0.40%	10	0.50%	10
LTR	8.10%	100	1.50%	50	0.00%	4	4.80%	70
DNA	2.80%	60	0.70%	20	5.30%	80	5.10%	80
Total	44.40%	170	3.10%	90	6.50%	90	10.50%	160

The complete genomes of fly, worm, and chromosomes 2 and 4 of mustard weed (as deposited at ncbi.nlm.nih.gov/genbank/genomes) were screened against the repeats in RepBase Update 5.02 (September 2000) with RepeatMasker at sensitive settings.

have suggested that small deletions occur at a rate that is 75-fold higher in flies than in mammals; the half-life of such nonfunctional DNA is estimated at 12 Myr for flies and 800 Myr for mammals¹⁶⁷. The rate of large deletions has not been systematically compared, but seems likely also to differ markedly.

(3) Whereas in the human two repeat families (LINE1 and Alu) account for 60% of all interspersed repeat sequence, the other organisms have no dominant families. Instead, the worm, fly and mustard weed genomes all contain many transposon families, each consisting of typically hundreds to thousands of elements. This difference may be explained by the observation that the vertically transmitted, long-term residential LINE and SINE elements represent 75% of interspersed repeats in the human genome, but only 5–25% in the other genomes. In contrast, the horizontally transmitted and shorter-lived DNA transposons represent only a small portion of all interspersed repeats in humans (6%) but a much larger fraction in fly, mustard weed and worm (25%, 49% and 87%, respectively). These features of the human genome are probably general to all mammals. The relative lack of horizontally transmitted elements may have its origin in the well developed immune system of mammals, as horizontal transfer requires infectious vectors, such as viruses, against which the immune system guards.

We also looked for differences among mammals, by comparing the transposons in the human and mouse genomes. As with the human genome, care is required in calibrating the substitution clock for the mouse genome. There is considerable evidence that the rate of substitution per Myr is higher in rodent lineages than in the hominid lineages^{139,168,169}. In fact, we found clear evidence for different rates of substitution by examining families of transposable elements whose insertions predate the divergence of the human and mouse lineages. In an analysis of 22 such families, we found that the substitution level was an average of 1.7-fold higher in mouse than human (not shown). (This is likely to be an underestimate because of an ascertainment bias against the most diverged copies.) The faster clock in mouse is also evident from the fact that the ancient LINE2 and MIR elements, which transposed before the mammalian radiation and are readily detectable in the human genome, cannot be readily identified in available mouse genomic sequence (Fig. 18).

We used the best available estimates to calibrate substitution levels and time¹⁶⁹. The ratio of substitution rates varied from about 1.7-fold higher over the past 100 Myr to about 2.6-fold higher over the past 25 Myr.

The analysis shows that, although the overall density of the four transposon types in human and mouse is similar, the age distribution is strikingly different (Fig. 18). Transposon activity in the mouse genome has not undergone the decline seen in humans and proceeds at a much higher rate. In contrast to their possible extinction in humans, LTR retrotransposons are alive and well in the

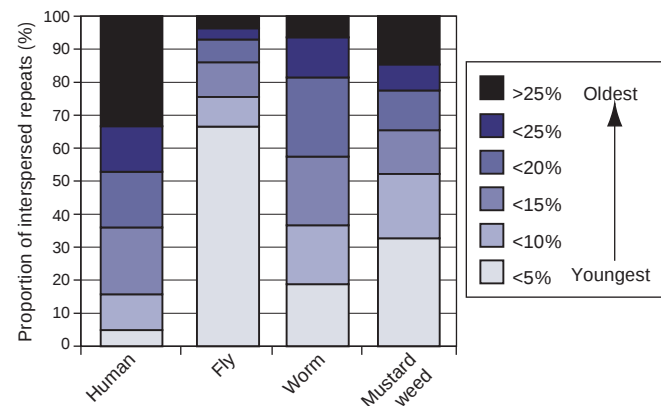


Figure 20 Comparison of the age of interspersed repeats in eukaryotic genomes. The copies of repeats were pooled by their nucleotide substitution level from the consensus.

mouse with such representatives as the active IAP family and putatively active members of the long-lived ERVL and MaLR families. LINE1 and a variety of SINEs are quite active. These evolutionary findings are consistent with the empirical observations that new spontaneous mutations are 30 times more likely to be caused by LINE insertions in mouse than in human (~3% versus 0.1%)¹⁷⁰ and 60 times more likely to be caused by transposable elements in general. It is estimated that around 1 in 600 mutations in human are due to transpositions, whereas 10% of mutations in mouse are due to transpositions (mostly IAP insertions).

The contrast between human and mouse suggests that the explanation for the decline of transposon activity in humans may lie in some fundamental difference between hominids and rodents. Population structure and dynamics would seem to be likely suspects. Rodents tend to have large populations, whereas hominid populations tend to be small and may undergo frequent bottlenecks. Evolutionary forces affected by such factors include inbreeding and genetic drift, which might affect the persistence of active transposable elements¹⁷¹. Studies in additional mammalian lineages may shed light on the forces responsible for the differences in the activity of transposable elements¹⁷².

Variation in the distribution of repeats. We next explored variation in the distribution of repeats across the draft genome sequence, by calculating the repeat density in windows of various sizes across the genome. There is striking variation at smaller scales.

Some regions of the genome are extraordinarily dense in repeats. The prizewinner appears to be a 525-kb region on chromosome Xp11, with an overall transposable element density of 89%. This region contains a 200-kb segment with 98% density, as well as a segment of 100 kb in which LINE1 sequences alone comprise 89% of the sequence. In addition, there are regions of more than 100 kb with extremely high densities of Alu (> 56% at three loci, including one on 7q11 with a 50-kb stretch of > 61% Alu) and the ancient transposons MIR (> 15% on chromosome 1p36) and LINE2 (> 18% on chromosome 22q12).

In contrast, some genomic regions are nearly devoid of repeats. The absence of repeats may be a sign of large-scale *cis*-regulatory elements that cannot tolerate being interrupted by insertions. The four regions with the lowest density of interspersed repeats in the human genome are the four homeobox gene clusters, HOXA, HOXB, HOXC and HOXD (Fig. 21). Each locus contains regions of around 100 kb containing less than 2% interspersed repeats. Ongoing sequence analysis of the four HOX clusters in mouse, rat and baboon shows a similar absence of transposable elements, and reveals a high density of conserved noncoding elements (K. Dewar and B. Birren, manuscript in preparation). The presence of a complex collection of regulatory regions may explain why individual HOX genes carried in transgenic mice fail to show proper regulation.

It may be worth investigating other repeat-poor regions, such as a region on chromosome 8q21 (1.5% repeat over 63 kb) containing a gene encoding a homeodomain zinc-finger protein (homologous to mouse pID 9663936), a region on chromosome 1p36 (5% repeat over 100 kb) with no obvious genes and a region on chromosome 18q22 (4% over 100 kb) containing three genes of unknown function (among which is KIAA0450). It will be interesting to see whether the homologous regions in the mouse genome have

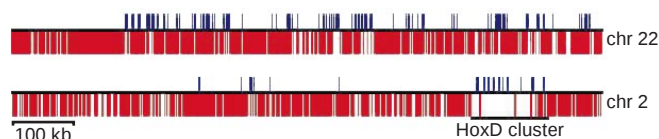


Figure 21 Two regions of about 1 Mb on chromosomes 2 and 22. Red bars, interspersed repeats; blue bars, exons of known genes. Note the deficit of repeats in the HoxD cluster, which contains a collection of genes with complex, interrelated regulation.

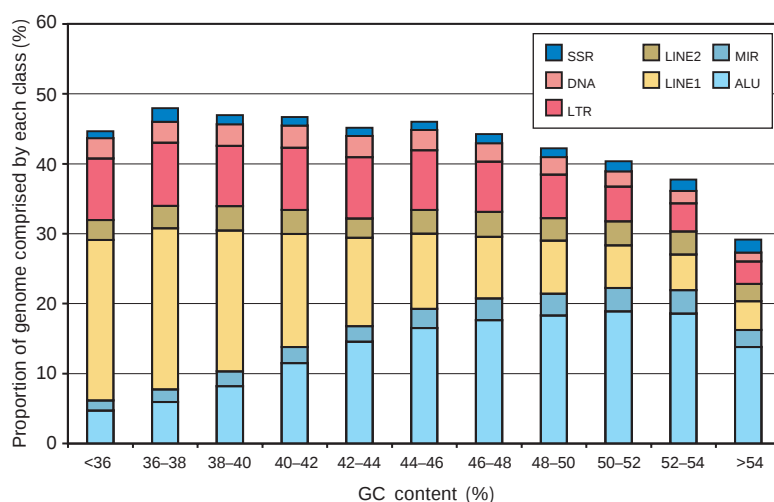


Figure 22 Density of the major repeat classes as a function of local GC content, in windows of 50 kb.

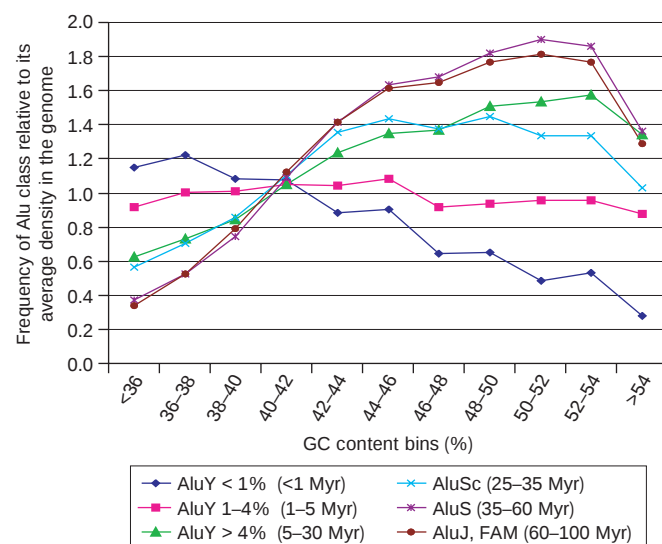


Figure 23 Alu elements target AT-rich DNA, but accumulate in GC-rich DNA. This graph shows the relative distribution of various Alu cohorts as a function of local GC content. The divergence levels (including CpG sites) and ages of the cohorts are shown in the key.

similarly resisted the insertion of transposable elements during rodent evolution.

Distribution by GC content. We next focused on the correlation between the nature of the transposons in a region and its GC content. We calculated the density of each repeat type as a function of the GC content in 50-kb windows (Fig. 22). As has been reported^{142,173–176}, LINE sequences occur at much higher density in AT-rich regions (roughly fourfold enriched), whereas SINEs (MIR, Alu) show the opposite trend (for Alu, up to fivefold lower in AT-rich DNA). LTR retrotransposons and DNA transposons show a more uniform distribution, dipping only in the most GC-rich regions.

The preference of LINES for AT-rich DNA seems like a reasonable way for a genomic parasite to accommodate its host, by targeting gene-poor AT-rich DNA and thereby imposing a lower mutational burden. Mechanistically, selective targeting is nicely explained by the fact that the preferred cleavage site of the LINE endonuclease is TTTT/A (where the slash indicates the point of cleavage), which is used to prime reverse transcription from the poly(A) tail of LINE RNA¹⁷⁷.

The contrary behaviour of SINEs, however, is baffling. How do

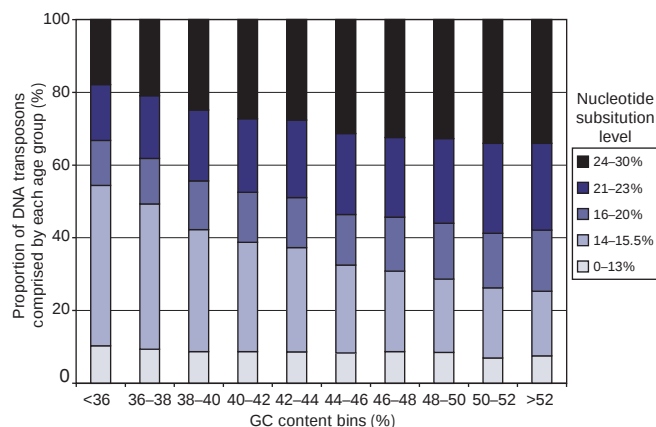


Figure 24 DNA transposon copies in AT-rich DNA tend to be younger than those in more GC-rich DNA. DNA transposon families were grouped into five age categories by their median substitution level (see Fig. 19). The proportion attributed to each age class is shown as a function of GC content. Similar patterns are seen for LINE1 and LTR elements.

SINEs accumulate in GC-rich DNA, particularly if they depend on the LINE transposition machinery¹⁷⁸. Notably, the same pattern is seen for the Alu-like B1 and the tRNA-derived SINEs in mouse and for MIR in human¹⁴². One possibility is that SINEs somehow target GC-rich DNA for insertion. The alternative is that SINEs initially insert with the same proclivity for AT-rich DNA as LINES, but that the distribution is subsequently reshaped by evolutionary forces^{142,179}.

We used the draft genome sequence to investigate this mystery by comparing the proclivities of young, adolescent, middle-aged and old Alus (Fig. 23). Strikingly, recent Alus show a preference for AT-rich DNA resembling that of LINES, whereas progressively older Alus show a progressively stronger bias towards GC-rich DNA. These results indicate that the GC bias must result from strong pressure: Fig. 23 shows that a 13-fold enrichment of Alus in GC-rich DNA has occurred within the last 30 Myr, and possibly more recently.

These results raise a new mystery. What is the force that produces the great and rapid enrichment of Alus in GC-rich DNA? One explanation may be that deletions are more readily tolerated in gene-poor AT-rich regions than in gene-rich GC-rich regions, resulting in older elements being enriched in GC-rich regions. Such an enrichment is seen for transposable elements such as

DNA transposons (Fig. 24). However, this effect seems too slow and too small to account for the observed remodelling of the Alu distribution. This can be seen by performing a similar analysis for LINE elements (Fig. 25). There is no significant change in the LINE distribution over the past 100 Myr, in contrast to the rapid change seen for Alu. There is an eventual shift after more than 100 Myr, although its magnitude is still smaller than seen for Alu.

These observations indicate that there may be some force acting particularly on Alus. This could be a higher rate of random loss of Alus in AT-rich DNA, negative selection against Alus in AT-rich DNA or positive selection in favour of Alus in GC-rich DNA. The first two possibilities seem unlikely because AT-rich DNA is

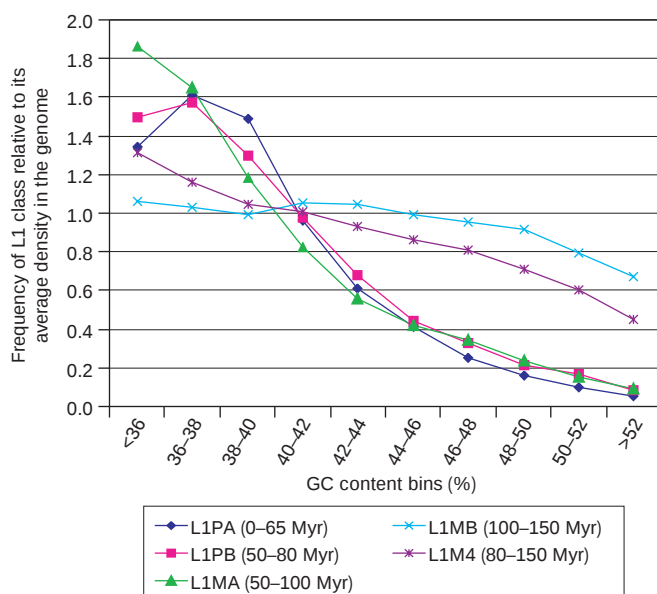


Figure 25 Distribution of various LINE cohorts as a function of local GC content. The divergence levels and ages of the cohorts are shown in the key. (The divergence levels were measured for the 3' UTR of the LINE1 element only, which is best characterized evolutionarily. This region contains almost no CpG sites, and thus 1% divergence level corresponds to a much longer time than for CpG-rich Alu copies).

gene-poor and tolerates the accumulation of other transposable elements. The third seems more feasible, in that it involves selecting in favour of the minority of Alus in GC-rich regions rather than against the majority that lie in AT-rich regions. But positive selection for Alus in GC-rich regions would imply that they benefit the organism.

Schmid¹⁸⁰ has proposed such a function for SINEs. This hypothesis is based on the observation that in many species SINEs are transcribed under conditions of stress, and the resulting RNAs specifically bind a particular protein kinase (PKR) and block its ability to inhibit protein translation¹⁸¹⁻¹⁸³. SINE RNAs would thus promote protein translation under stress. SINE RNA may be well suited to such a role in regulating protein translation, because it can be quickly transcribed in large quantities from thousands of elements and it can function without protein translation. Under this theory, there could be positive selection for SINEs in readily transcribed open chromatin such as is found near genes. This could explain the retention of Alus in gene-rich GC-rich regions. It is also consistent with the observation that SINE density in AT-rich DNA is higher near genes¹⁴².

Further insight about Alus comes from the relationship between Alu density and GC content on individual chromosomes (Fig. 26). There are two outliers. Chromosome 19 is even richer in Alus than predicted by its (high) GC content; the chromosome comprises 2% of the genome, but contains 5% of Alus. On the other hand, chromosome Y shows the lowest density of Alus relative to its GC content, being higher than average for GC content less than 40% and lower than average for GC content over 40%. Even in AT-rich DNA, Alus are under-represented on chromosome Y compared with other young interspersed repeats (see below). These phenomena may be related to an unusually high gene density on chromosome 19 and an unusually low density of somatically active genes on chromosome Y (both relative to GC content). This would be consistent with the idea that Alu correlates not with GC content but with actively transcribed genes.

Our results may support the controversial idea that SINEs actually earn their keep in the genome. Clearly, much additional work will be needed to prove or disprove the hypothesis that SINEs are genomic symbionts.

Biases in human mutation. Indirect studies have suggested that nucleotide substitution is not uniform across mammalian

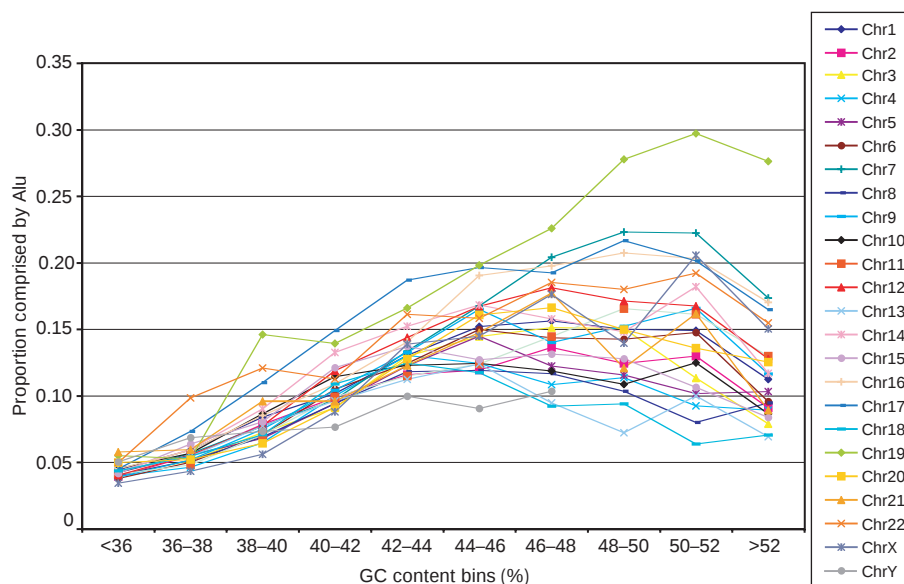


Figure 26 Comparison of the Alu density of each chromosome as a function of local GC content. At higher GC levels, the Alu density varies widely between chromosomes, with chromosome 19 being a particular outlier. In contrast, the LINE1 density pattern is quite

uniform for most chromosomes, with the exception of a 1.5 to 2-fold over-representation in AT-rich regions of the X and Y chromosomes (not shown).

genomes^{184–187}. By studying sets of repeat elements belonging to a common cohort, one can directly measure nucleotide substitution rates in different regions of the genome. We find strong evidence that the pattern of neutral substitution differs as a function of local GC content (Fig. 27). Because the results are observed in repetitive elements throughout the genome, the variation in the pattern of nucleotide substitution seems likely to be due to differences in the underlying mutational process rather than to selection.

The effect can be seen most clearly by focusing on the substitution process $\gamma \leftrightarrow \alpha$, where γ denotes GC or CG base pairs and α denotes AT or TA base pairs. If K is the equilibrium constant in the direction of α base pairs (defined by the ratio of the forward and reverse rates), then the equilibrium GC content should be $1/(1 + K)$. Two observations emerge.

First, there is a regional bias in substitution patterns. The equilibrium constant varies as a function of local GC content: γ base pairs are more likely to mutate towards α base pairs in AT-rich regions than in GC-rich regions. For the analysis in Fig. 27, the equilibrium constant K is 2.5, 1.9 and 1.2 when the draft genome sequence is partitioned into three bins with average GC content of 37, 43 and 50%, respectively. This bias could be due to a reported tendency for GC-rich regions to replicate earlier in the cell cycle than AT-rich regions and for guanine pools, which are limiting for

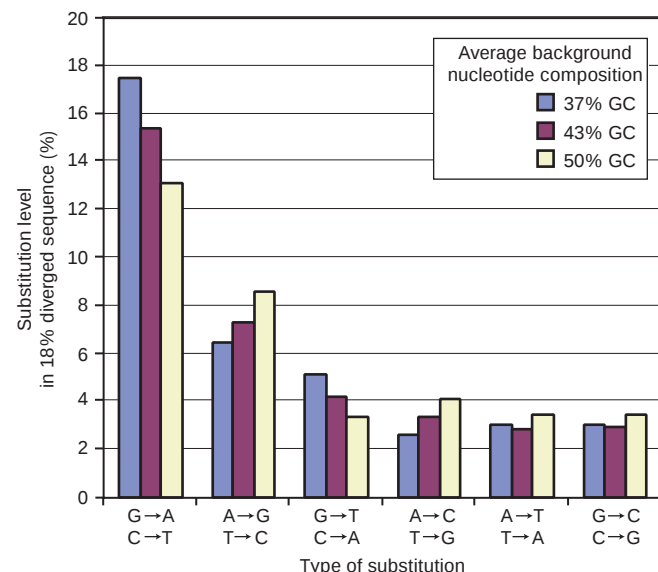


Figure 27 Substitution patterns in interspersed repeats differ as a function of GC content. We collected all copies of five DNA transposons (Tigger1, Tigger2, Charlie3, MER1 and HSMAR2), chosen for their high copy number and well defined consensus sequences. DNA transposons are optimal for the study of neutral substitutions: they do not segregate into subfamilies with diagnostic differences, presumably because they are short-lived and new active families do not evolve in a genome (see text). Duplicates and close paralogues resulting from duplication after transposition were eliminated. The copies were grouped on the basis of GC content of the flanking 1,000 bp on both sides and aligned to the consensus sequence (representing the state of the copy at integration). Recursive efforts using parameters arising from this study did not change the alignments significantly. Alignments were inspected by hand, and obvious misalignments caused by insertions and duplications were eliminated. Substitutions ($n = 80,000$) were counted for each position in the consensus, excluding those in CpG dinucleotides, and a substitution frequency matrix was defined. From the matrices for each repeat (which corresponded to different ages), a single rate matrix was calculated for these bins of GC content ($< 40\%$ GC, $40–47\%$ GC and $> 47\%$ GC). Data are shown for a repeat with an average divergence (in non-CpG sites) of 18% in 43% GC content (the repeat has slightly higher divergence in AT-rich DNA and lower in GC-rich DNA). From the rate matrix, we calculated log-likelihood matrices with different entropies (divergence levels), which are theoretically optimal for alignments of neutrally diverged copies to their common ancestral state (A. Kas and A. F. A. Smit, unpublished). These matrices are in use by the RepeatMasker program.

DNA replication, to become depleted late in the cell cycle, thereby resulting in a small but significant shift in substitution towards α base pairs^{186,188}. Another theory proposes that many substitutions are due to differences in DNA repair mechanisms, possibly related to transcriptional activity and thereby to gene density and GC content^{185,189,190}.

There is also an absolute bias in substitution patterns resulting in directional pressure towards lower GC content throughout the human genome. The genome is not at equilibrium with respect to the pattern of nucleotide substitution: the expected equilibrium GC content corresponding to the values of K above is 29, 35 and 44% for regions with average GC contents of 37, 43 and 50%, respectively. Recent observations on SNPs¹⁹⁰ confirm that the mutation pattern in GC-rich DNA is biased towards α base pairs; it should be possible to perform similar analyses throughout the genome with the availability of 1.4 million SNPs^{97,191}. On the basis solely of nucleotide substitution patterns, the GC content would be expected to be about 7% lower throughout the genome.

What accounts for the higher GC content? One possible explanation is that in GC-rich regions, a considerable fraction of the nucleotides is likely to be under functional constraint owing to the high gene density. Selection on coding regions and regulatory CpG islands may maintain the higher-than-predicted GC content. Another is that throughout the rest of the genome, a constant influx of transposable elements tends to increase GC content (Fig. 28). Young repeat elements clearly have a higher GC content than their surrounding regions, except in extremely GC-rich regions. Moreover, repeat elements clearly shift with age towards a lower GC content, closer to that of the neighbourhood in which they reside. Much of the 'non-repeat' DNA in AT-rich regions probably consists of ancient repeats that are not detectable by current methods and that have had more time to approach the local equilibrium value.

The repeats can also be used to study how the mutation process is affected by the immediately adjacent nucleotide. Such 'context effects' will be discussed elsewhere (A. Kas and A. F. A. Smit, unpublished results).

Fast living on chromosome Y. The pattern of interspersed repeats can be used to shed light on the unusual evolutionary history of chromosome Y. Our analysis shows that the genetic material on

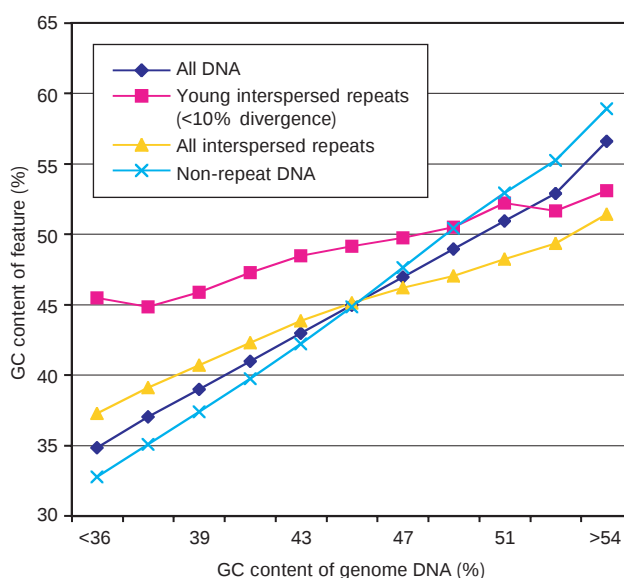


Figure 28 Interspersed repeats tend to diminish the differences between GC bins, despite the fact that GC-rich transposable elements (specifically Alu) accumulate in GC-rich DNA, and AT-rich elements (LINE1) in AT-rich DNA. The GC content of particular components of the sequence (repeats, young repeats and non-repeat sequence) was calculated as a function of overall GC content.

chromosome Y is unusually young, probably owing to a high tolerance for gain of new material by insertion and loss of old material by deletion. Several lines of evidence support this picture. For example, LINE elements on chromosome Y are on average much younger than those on autosomes (not shown). Similarly, MaLR-family retrotransposons on chromosome Y are younger than those on autosomes, with the representation of subfamilies showing a strong inverse correlation with the age of the subfamily. Moreover, chromosome Y has a relative over-representation of the younger retroviral class II (ERV_K) and a relative under-representation of the primarily older class III (ERV_L) compared with other chromosomes. Overall, chromosome Y seems to maintain a youthful appearance by rapid turnover.

Interspersed repeats on chromosome Y can also be used to estimate the relative mutation rates, α_m and α_f , in the male and female germ lines. Chromosome Y always resides in males, whereas chromosome X resides in females twice as often as in males. The substitution rates, μ_Y and μ_X , on these two chromosomes should thus be in the ratio $\mu_Y:\mu_X = (\alpha_m):(\alpha_m + 2\alpha_f)/3$, provided that one considers equivalent neutral sequences. Several authors have estimated the mutation rate in the male germline to be fivefold higher than in the female germline, by comparing the rates of evolution of X- and Y-linked genes in humans and primates. However, Page and colleagues¹⁹² have challenged these estimates as too high. They studied a 39-kb region that is apparently devoid of genes and resides within a large segmental duplication from X to Y that occurred 3–4 Myr ago in the human lineage. On the basis of phylogenetic analysis of the sequence on human Y and human, chimp and gorilla X, they obtained a much lower estimate of $\mu_Y:\mu_X = 1.36$, corresponding to $\alpha_m:\alpha_f = 1.7$. They suggested that the other estimates may have been higher because they were based on much longer evolutionary periods or because the genes studied may have been under selection.

Our database of human repeats provides a powerful resource for addressing this question. We identified the repeat elements from recent subfamilies (effectively, birth cohorts dating from the past 50 Myr) and measured the substitution rates for subfamily members on chromosomes X and Y (Fig. 29). There is a clear linear relationship with a slope of $\mu_Y:\mu_X = 1.57$ corresponding to $\alpha_m:\alpha_f = 2.1$. The estimate is in reasonable agreement with that of Page *et al.*, although it is based on much more total sequence (360 kb on Y, 1.6 Mb on X) and a much longer time period. In particular, the discrepancy with earlier reports is not explained by recent changes in the human lineage. Various theories have been proposed for the higher mutation rate in the male germline, including the greater number of cell divisions in the formation of sperm than eggs and different repair mechanisms in sperm and eggs.

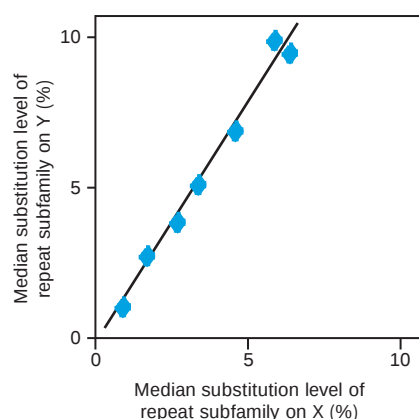


Figure 29 Higher substitution rate on chromosome Y than on chromosome X. We calculated the median substitution level (excluding CpG sites) for copies of the most recent L1 subfamilies (L1Hs–L1PA8) on the X and Y chromosomes. Only the 3' UTR of the L1 element was considered because its consensus sequence is best established.

Active transposons. We were interested in identifying the youngest retrotransposons in the draft genome sequence. This set should contain the currently active retrotransposons, as well as the insertion sites that are still polymorphic in the human population.

The youngest branch in the phylogenetic tree of human LINE1 elements is called L1Hs (ref. 158); it differs in its 3' untranslated region (UTR) by 12 diagnostic substitutions from the next oldest subfamily (L1PA2). Within the L1Hs family, there are two subsets referred to as Ta and pre-Ta, defined by a diagnostic trinucleotide^{193,194}. All active L1 elements are thought to belong to these two subsets, because they account for all 14 known cases of human disease arising from new L1 transposition (with 13 belonging to the Ta subset and one to the pre-Ta subset)^{195,196}. These subsets are also of great interest for population genetics because at least 50% are still segregating as polymorphisms in the human population^{194,197}; they provide powerful markers for tracing population history because they represent unique (non-recurrent and non-reversible) genetic events that can be used (along with similarly polymorphic Alus) for reconstructing human migrations.

LINE1 elements that are retrotransposition-competent should consist of a full-length sequence and should have both ORFs intact. Eleven such elements from the Ta subset have been identified, including the likely progenitors of mutagenic insertions into the factor VIII and dystrophin genes^{198–202}. A cultured cell retrotransposition assay has revealed that eight of these elements remain retrotransposition-competent^{200,202,203}.

We searched the draft genome sequence and identified 535 LINEs belonging to the Ta subset and 415 belonging to the pre-Ta subset. These elements provide a large collection of tools for probing human population history. We also identified those consisting of full-length elements with intact ORFs, which are candidate active LINEs. We found 39 such elements belonging to the Ta subset and 22 belonging to the pre-Ta subset; this substantially increases the number in the first category and provides the first known examples in the second category. These elements can now be tested for retrotransposition competence in the cell culture assay. Preliminary analysis resulted in the identification of two of these elements as the likely progenitors of mutagenic insertions into the β -globin and RP2 genes (R. Badge and J. V. Moran, unpublished data). Similar analyses should allow the identification of the progenitors of most, if not all, other known mutagenic L1 insertions.

L1 elements can carry extra DNA if transcription extends through the native transcriptional termination site into flanking genomic DNA. This process, termed L1-mediated transduction, provides a means for the mobilization of DNA sequences around the genome and may be a mechanism for 'exon shuffling'²⁰⁴. Twenty-one per cent of the 71 full-length L1s analysed contained non-L1-derived sequences before the 3' target-site duplication site, in cases in which the site was unambiguously recognizable. The length of the transduced sequence was 30–970 bp, supporting the suggestion that 0.5–1.0% of the human genome may have arisen by LINE-based transduction of 3' flanking sequences^{205,206}.

Our analysis also turned up two instances of 5' transduction (145 bp and 215 bp). Although this possibility had been suggested on the basis of cell culture models^{195,203}, these are the first documented examples. Such events may arise from transcription initiating in a cellular promoter upstream of the L1 elements. L1 transcription is generally confined to the germline^{207,208}, but transcription from other promoters could explain a somatic L1 retrotransposition event that resulted in colon cancer²⁰⁶.

Transposons as a creative force. The primary force for the origin and expansion of most transposons has been selection for their ability to create progeny, and not a selective advantage for the host. However, these selfish pieces of DNA have been responsible for important innovations in many genomes, for example by contributing regulatory elements and even new genes.

Twenty human genes have been recognized as probably derived

from transposons^{142,209}. These include the RAG1 and RAG2 recombinases and the major centromere-binding protein CENPB. We scanned the draft genome sequence and identified another 27 cases, bringing the total to 47 (Table 13; refs 142, 209). All but four are derived from DNA transposons, which give rise to only a small proportion of the interspersed repeats in the genome. Why there are so many DNA transposase-like genes, many of which still contain the critical residues for transposase activity, is a mystery.

To illustrate this concept, we describe the discovery of one of the new examples. We searched the draft genome sequence to identify the autonomous DNA transposon responsible for the distribution of the non-autonomous MER85 element, one of the most recently (40–50 Myr ago) active DNA transposons. Most non-autonomous elements are internal deletion products of a DNA transposon. We identified one instance of a large (1,782 bp) ORF flanked by the 5' and 3' halves of a MER85 element. The ORF encodes a novel protein (partially published as pID 6453533) whose closest homologue is the transposase of the piggyBac DNA transposon, which is found in insects and has the same characteristic TTAA target-site duplications²¹⁰ as MER85. The ORF is actively transcribed in fetal brain and in cancer cells. That it has not been lost to mutation in 40–50 Myr of evolution (whereas the flanking, noncoding, MER85-

like termini show the typical divergence level of such elements) and is actively transcribed provides strong evidence that it has been adopted by the human genome as a gene. Its function is unknown.

LINE1 activity clearly has also had fringe benefits. We mentioned above the possibility of exon reshuffling by cotranscription of neighbouring DNA. The LINE1 machinery can also cause reverse transcription of genic mRNAs, which typically results in nonfunctional processed pseudogenes but can, occasionally, give rise to functional processed genes. There are at least eight human and eight mouse genes for which evidence strongly supports such an origin²¹¹ (see <http://www-ifi.uni-muenster.de/exapted-retrogenes/tables.html>). Many other intronless genes may have been created in the same way.

Transposons have made other creative contributions to the genome. A few hundred genes, for example, use transcriptional terminators donated by LTR retrotransposons (data not shown). Other genes employ regulatory elements derived from repeat elements²¹¹.

Simple sequence repeats

Simple sequence repeats (SSRs) are a rather different type of repetitive structure that is common in the human genome—perfect or slightly imperfect tandem repeats of a particular *k*-mer. SSRs with a short repeat unit (*n* = 1–13 bases) are often termed microsa-

Table 13 Human genes derived from transposable elements

GenBank ID*	Gene name	Related transposon family†	Possible fusion gene§	Newly recognized derivation
nID 3150436	BC200	FLAM Alu‡		
pID 2330017	Telomerase	non-LTR retrotransposon		
pID 1196425	HERV-3 env	Retroviridae/HERV-R‡		
pID 4773880	Syncytin	Retroviridae/HERV-W‡		
pID 131827	RAG1 and 2	Tc1-like		
pID 29863	CENP-B	Tc1/Pogo		
EST 2529718		Tc1/Pogo		+
PID 10047247		Tc1/Pogo/Pogo		+
EST 4524463		Tc1/Pogo/Pogo		+
pID 4504807	Jerky	Tc1/Pogo/Tigger		
pID 7513096	JRKL	Tc1/Pogo/Tigger		
EST 5112721		Tc1/Pogo/Tigger		+
EST 11097233		Tc1/Pogo/Tigger		+
EST 6986275	Sancho	Tc1/Pogo/Tigger		
EST 8616450		Tc1/Pogo/Tigger		+
EST 8750408		Tc1/Pogo/Tigger		+
EST 5177004		Tc1/Pogo/Tigger		+
PID 3413884	KIAA0461	Tc1/Pogo/Tc2	+	
PID 7959287	KIAA1513	Tc1/Pogo/Tc2		+
PID 2231380		Tc1/Mariner/Hsmar1‡	+	
EST 10219887		hAT/Hobo	+	+
PID 6581095	Buster1	hAT/Charlie	+	
PID 7243087	Buster2	hAT/Charlie	+	
PID 6581097	Buster3	hAT/Charlie		
PID 7662294	KIAA0766	hAT/Charlie	+	
PID 10439678		hAT/Charlie		+
PID 7243087	KIAA1353	hAT/Charlie		+
PID 7021900		hAT/Charlie/Charlie3‡		+
PID 4263748		hAT/Charlie/Charlie8‡	+	
EST 8161741		hAT/Charlie/Charlie9‡		+
pID 4758872	DAP4, pP52 ^{IPK}	hAT/Tip100/Zaphod		
EST 10990063		hAT/Tip100/Zaphod		+
EST 10101591		hAT/Tip100/Zaphod		+
pID 7513011	KIAA0543	hAT/Tip100/Tip100	+	
pID 10439744		hAT/Tip100/Tip100		+
pID 10047247	KIAA1586	hAT/Tip100/Tip100		+
pID 10439762		hAT/Tip100	+	+
EST 10459804		hAT/Tip100		+
pID 4160548	Tramp	hAT/Tam3		+
BAC 3522927		hAT/Tam3		+
pID 3327088	KIAA0637	hAT/Tam3	+	
EST 1928552		hAT/Tam3		+
pID 6453533		piggyBac/MER85‡	+	
EST 3594004		piggyBac/MER85‡		+
BAC 4309921		piggyBac/MER85‡		+
EST 4073914		piggyBac/MER75‡		+
EST 1963278		piggyBac		+

The Table lists 47 human genes, with a likely origin in up to 38 different transposon copies.

* Where available, the GenBank ID numbers are given for proteins, otherwise a representative EST or a clone name is shown. Six groups (two or three genes each) have similarity at the DNA level well beyond that observed between different DNA transposon families in the genome; they are indicated in italics, with all but the initial member of each group indented. This could be explained if the genes were paralogous (derived from a single inserted transposon and subsequently duplicated).

† Classification of the transposon.

‡ Indicates that the transposon from which the gene is derived is precisely known.

§ Proteins probably formed by fusion of a cellular and transposon gene; many have acquired zinc-finger domains.

|| Not previously reported as being derived from transposable element genes. The remaining genes can be found in refs 142, 209.

Table 14 SSR content of the human genome

Length of repeat unit	Average bases per Mb	Average number of SSR elements per Mb
1	1,660	36.7
2	5,046	43.1
3	1,013	11.8
4	3,383	32.5
5	2,686	17.6
6	1,376	15.2
7	906	8.4
8	1,139	11.1
9	900	8.6
10	1,576	8.6
11	770	8.7

SSRs were identified by using the computer program Tandem Repeat Finder with the following parameters: match score 2, mismatch score 3, indel 5, minimum alignment 50, maximum repeat length 500, minimum repeat length 1.

tellites, whereas those with longer repeat units ($n = 14$ –500 bases) are often termed minisatellites. With the exception of poly(A) tails from reverse transcribed messages, SSRs are thought to arise by slippage during DNA replication^{212,213}.

We compiled a catalogue of all SSRs over a given length in the human draft genome sequence, and studied their properties (Table 14). SSRs comprise about 3% of the human genome, with the greatest single contribution coming from dinucleotide repeats (0.5%). (The precise criteria for the number of repeat units and the extent of divergence allowed in an SSR affect the exact census, but not the qualitative conclusions.)

There is approximately one SSR per 2 kb (the number of non-overlapping tandem repeats is 437 per Mb). The catalogue confirms various properties of SSRs that have been inferred from sampling approaches (Table 15). The most frequent dinucleotide repeats are AC and AT (50 and 35% of dinucleotide repeats, respectively), whereas AG repeats (15%) are less frequent and GC repeats (0.1%) are greatly under-represented. The most frequent trinucleotides are AAT and AAC (33% and 21%, respectively), whereas ACC (4.0%), AGC (2.2%), ACT (1.4%) and ACG (0.1%) are relatively rare. Overall, trinucleotide SSRs are much less frequent than dinucleotide SSRs²¹⁴.

SSRs have been extremely important in human genetic studies, because they show a high degree of length polymorphism in the human population owing to frequent slippage by DNA polymerase during replication. Genetic markers based on SSRs—particularly (CA)_n repeats—have been the workhorse of most human disease-mapping studies^{101,102}. The availability of a comprehensive catalogue of SSRs is thus a boon for human genetic studies.

The SSR catalogue also allowed us to resolve a mystery regarding mammalian genetic maps. Such genetic maps in rat, mouse and human have a deficit of polymorphic (CA)_n repeats on chromosome X^{30,101}. There are two possible explanations for this deficit. There may simply be fewer (CA)_n repeats on chromosome X; or (CA)_n repeats may be as dense on chromosome X but less polymorphic in

Table 15 SSRs by repeat unit

Repeat unit	Number of SSRs per Mb
AC	27.7
AT	19.4
AG	8.2
GC	0.1
AAT	4.1
AAC	2.6
AGG	1.5
AAG	1.4
ATG	0.7
CGG	0.6
ACC	0.4
AGC	0.3
ACT	0.2
ACG	0.0

SSRs were identified as in Table 14.

the population. In fact, analysis of the draft genome sequence shows that chromosome X has the same density of (CA)_n repeats per Mb as the autosomes (data not shown). Thus, the deficit of polymorphic markers relative to autosomes results from population genetic forces. Possible explanations include that chromosome X has a smaller effective population size, experiences more frequent selective sweeps reducing diversity (owing to its hemizyosity in males), or has a lower mutation rate (owing to its more frequent passage through the less mutagenic female germline). The availability of the draft genome sequence should provide ways to test these alternative explanations.

Segmental duplications

A remarkable feature of the human genome is the segmental duplication of portions of genomic sequence^{215–217}. Such duplications involve the transfer of 1–200-kb blocks of genomic sequence to one or more locations in the genome. The locations of both donor and recipient regions of the genome are often not tandemly arranged, suggesting mechanisms other than unequal crossing-over for their origin. They are relatively recent, inasmuch as strong sequence identity is seen in both exons and introns (in contrast to regions that are considered to show evidence of ancient duplications, characterized by similarities only in coding regions). Indeed, many such duplications appear to have arisen in very recent evolutionary time, as judged by high sequence identity and by their absence in closely related species.

Segmental duplications can be divided into two categories. First, interchromosomal duplications are defined as segments that are duplicated among nonhomologous chromosomes. For example, a 9.5-kb genomic segment of the adrenoleukodystrophy locus from Xq28 has been duplicated to regions near the centromeres of chromosomes 2, 10, 16 and 22 (refs 218, 219). Anecdotal observations suggest that many interchromosomal duplications map near the centromeric and telomeric regions of human chromosomes^{218–233}.

The second category is intrachromosomal duplications, which occur within a particular chromosome or chromosomal arm. This category includes several duplicated segments, also known as low copy repeat sequences, that mediate recurrent chromosomal structural rearrangements associated with genetic disease^{215,217}. Examples on chromosome 17 include three copies of a roughly 200-kb repeat separated by around 5 Mb and two copies of a roughly 24-kb repeat separated by 1.5 Mb. The copies are so similar (99% identity) that paralogous recombination events can occur, giving rise to contiguous gene syndromes: Smith–Magenis syndrome and Charcot–Marie–Tooth syndrome 1A, respectively^{34,234}. Several other examples are known and are also suspected to be responsible for recurrent microdeletion syndromes (for example, Prader–Willi/Angelman,

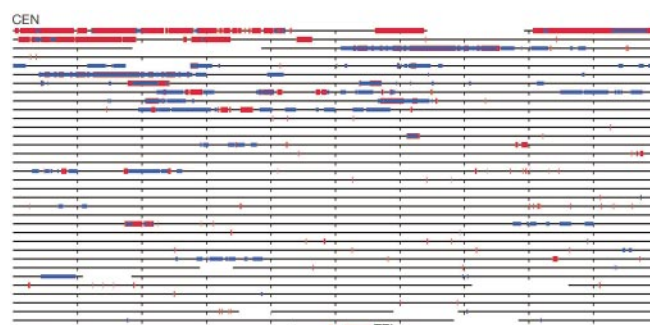


Figure 30 Duplication landscape of chromosome 22. The size and location of intrachromosomal (blue) and interchromosomal (red) duplications are depicted for chromosome 22q, using the PARASIGHT computer program (Bailey and Eichler, unpublished). Each horizontal line represents 1 Mb (ticks, 100-kb intervals). The chromosome sequence is oriented from centromere (top left) to telomere (bottom right). Pairwise alignments with >90% nucleotide identity and >1 kb long are shown. Gaps within the chromosomal sequence are of known size and shown as empty space.

velocardiofacial/DiGeorge and Williams' syndromes^{215,235–240}).

Until now, the identification and characterization of segmental duplications have been based on anecdotal reports—for example, finding that certain probes hybridize to multiple chromosomal sites or noticing duplicated sequence at certain recurrent chromosomal breakpoints. The availability of the entire genomic sequence will make it possible to explore the nature of segmental duplications more systematically. This analysis can begin with the current state of the draft genome sequence, although caution is required because some apparent duplications may arise from a failure to merge sequence contigs from overlapping clones. Alternatively, erroneous

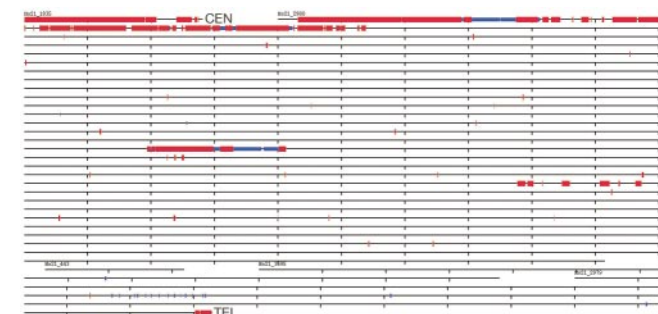


Figure 31 Duplication landscape of chromosome 21. The size and location of intrachromosomal (blue) and interchromosomal (red) duplications are depicted along the sequence of the long arm of chromosome 21. Gaps between finished sequence are denoted by empty space but do not represent actual gap size.

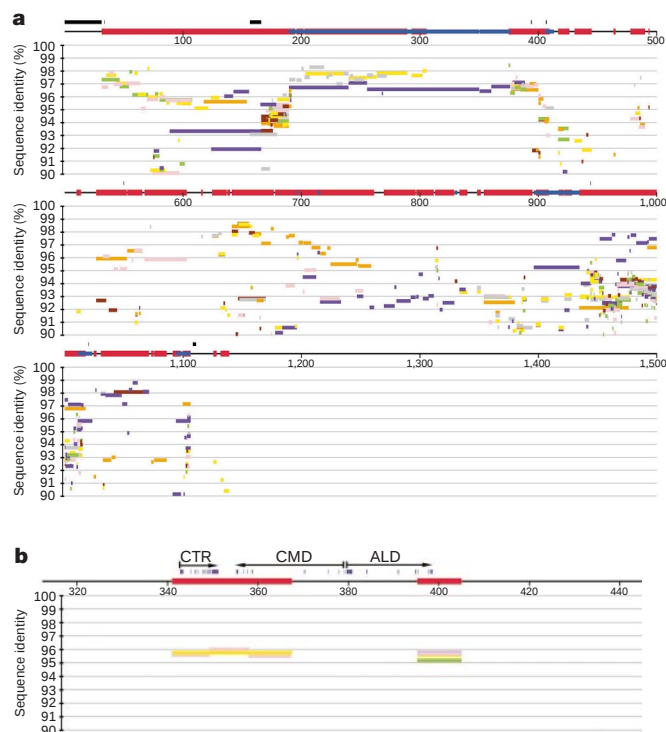


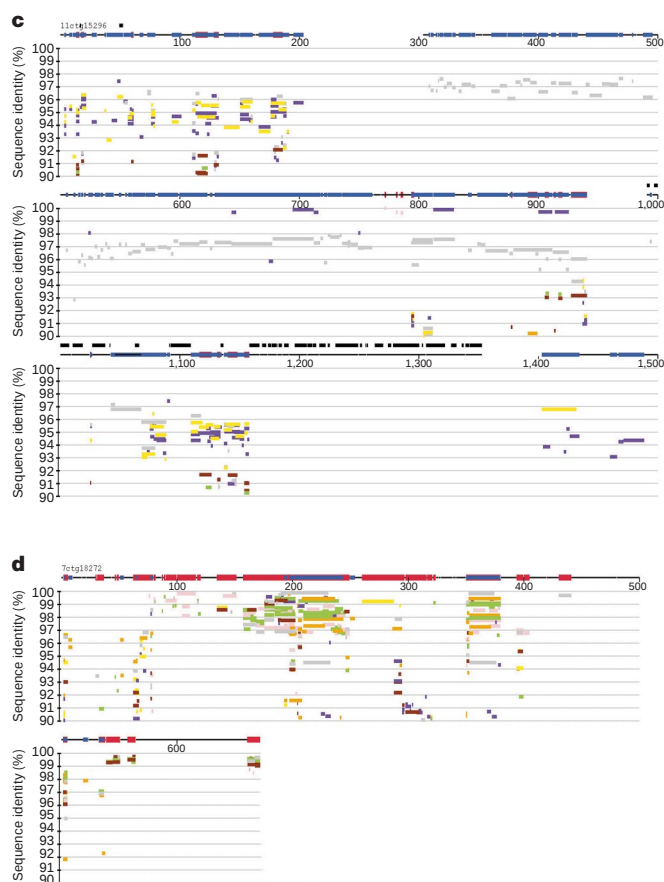
Figure 32 Mosaic patterns of duplications. Panels depict various patterns of duplication within the human genome (PARASIGHT). For each region, a segment of draft genome sequence (100–500 kb) is shown with both interchromosomal (red) and intrachromosomal (blue) duplications displayed along the horizontal line. Below the line, each separate

assembly of closely related sequences from nonoverlapping clones may underestimate the true frequency of such features, particularly among those segments with the highest sequence similarity. Accordingly, we adopted a conservative approach for estimating such duplication from the available draft genome sequence.

Pericentromeres and subtelomeres. We began by re-evaluating the finished sequences of chromosomes 21 and 22. The initial papers on these chromosomes^{93,94} noted some instances of interchromosomal duplication near each centromere. With the ability now to compare these chromosomes to the vast majority of the genome, it is apparent that the regions near the centromeres consist almost entirely of interchromosomal duplicated segments, with little or no unique sequence. Smaller regions of interchromosomal duplication are also observed near the telomeres.

Chromosome 22 contains a region of 1.5 Mb adjacent to the centromere in which 90% of sequence can now be recognized to consist of interchromosomal duplication (Fig. 30). Conversely, 52% of the interchromosomal duplications on chromosome 22 were located in this region, which comprises only 5% of the chromosome. Also, the subtelomeric end consists of a 50-kb region consisting almost entirely of interchromosomal duplications.

Chromosome 21 presents a similar landscape (Fig. 31). The first 1 Mb after the centromere is composed of interchromosomal repeats, as well as the largest (> 200 kb) block of intrachromosomally duplicated material. Again, most interchromosomal duplications on the chromosome map to this region and the most subtelomeric region (30 kb) shows extensive duplication among nonhomologous chromosomes.



sequence duplication is indicated (with a distinct colour) relative to per cent nucleotide identity for the duplicated segment (y axis). Black bars show the relative locations of large blocks of heterochromatic sequences (alpha, gamma and HSAT sequence). **a**, An active pericentromeric region on chromosome 21. **b**, An ancestral region from Xq28 that has contributed various 'genetic' segments to pericentromeric regions. **c**, A pericentromeric

The pericentromeric regions are structurally very complex, as illustrated for chromosome 21 in Fig. 32a. The pericentromeric regions appear to have been bombarded by successive insertions of duplications; the insertion events must be fairly recent because the degree of sequence conservation with the genomic source loci is fairly high (90–100%, with an apparent peak around 96%). Distinct insertions are typically separated by AT-rich or GC-rich minisatellite-like repeats that have been hypothesized to have a functional role in targeting duplications to these regions^{233,241}.

A single genomic source locus often gives rise to pericentromeric copies on multiple chromosomes, with each having essentially the same breakpoints and the same degree of divergence. An example of such a source locus on Xq28 is shown in Fig. 32b. Phylogenetic analysis has suggested a two-step mechanism for the origin and dispersal of these segments, whereby an initial segmental duplication in the pericentromeric region of one chromosome occurs and is then redistributed as part of a larger cassette to other such regions²⁴².

A comprehensive analysis for all chromosomes will have to await complete sequencing of the genome, but the evidence from the draft genome sequence indicates that the same picture is likely to be seen throughout the genome. Several papers have analysed finished segments within pericentromeric regions of chromosomes 2 (160 kb), 10 (400 kb) and 16 (300 kb), all of which show extensive interchromosomal segmental duplication^{215,219,232,233}. An example from another pericentromeric region on chromosome 11 is shown in Fig. 32c. Interchromosomal duplications in subtelomeric regions also appear to be a fairly general phenomenon, as illustrated by a large tract (~500 kb) of complex duplication on chromosome 7 (Fig. 32d).

The explanation for the clustering of segmental duplications may be that the genome has a damage-control mechanism whereby chromosomal breakage products are preferentially inserted into pericentromeric and, to a lesser extent, subtelomeric regions. The possibility of a specific mechanism for the insertion of these sequences has been suggested on the basis of the unusual sequences found flanking the insertions. Although it is also possible that these regions simply have greater tolerance for large insertions, many large gene-poor 'deserts' have been identified⁹³ and there is no accumulation of duplicated segments within these regions. Along with the fact that transitions between duplicons (from different regions of the genome) occur at specific sequences, this suggests that active recruitment of duplications to such regions may occur. In any case, the duplicated regions are in general young (with many duplications showing <6% nucleotide divergence from their source loci) and in constant flux, both through additional duplications and by large-scale exchange among similar chromosomal environments. There is evidence of structural polymorphism in the human population, such as the presence or absence of olfactory receptor segments located within the telomeric regions of several human chromosomes^{226,227}.

Genome-wide analysis of segmental duplications. We also performed a global genome-wide analysis to characterize the amount of segmental duplication in the genome. We 'repeat-masked' the known interspersed repeats in the draft genome sequence and compared the remaining draft genomic sequence with itself in a massive all-by-all BLASTN similarity search. We excluded matches in which the sequence identity was so high that it might reflect artefactual duplications resulting from a failure to overlap sequence contigs correctly in assembling the draft genome sequence. Specifically, we considered only matches with less than 99.5% identity for finished sequence and less than 98% identity for unfinished sequence.

We took several approaches to avoid counting artefactual duplications in the sequence. In the first approach, we studied only finished sequence. We compared the finished sequence with itself, to identify segments of at least 1 kb and 90–99.5% sequence identity. This analysis will underestimate the extent of segmental

duplication, because it requires that at least two copies of the segment are present in the finished sequence and because some true duplications have over 99.5% identity.

The finished sequence consists of at least 3.3% segmental duplication (Table 16). Interchromosomal duplication accounts for about 1.5% and intrachromosomal duplication for about 2%, with some overlap (0.2%) between these categories. We analysed the lengths and divergence of the segmental duplications (Fig. 33). The duplications tend to be large (10–50 kb) and highly homologous, especially for the interchromosomal segments. The sequence divergence for the interchromosomal duplications appears to peak between 96.5% and 97.5%. This may indicate that interchromosomal duplications occurred in a punctuated manner. It will be intriguing to investigate whether such genomic upheaval has a role in speciation events.

In a second approach, we compared the entire human draft genome sequence (finished and unfinished) with itself to identify duplications with 90–98% sequence identity (Table 17). The draft genome sequence contains at least 3.6% segmental duplication. The actual proportion will be significantly higher, because we excluded many true matches with more than 98% sequence identity (at least 1.1% of the finished sequence). Although exact measurement must await a finished sequence, the human genome seems likely to contain about 5% segmental duplication, with most of this sequence in large blocks (>10 kb). Such a high proportion of large duplications clearly distinguishes the human genome from other sequenced genomes, such as the fly and worm (Table 18).

The structure of large highly paralogous regions presents one of the 'serious and unanticipated challenges' to producing a finished sequence of the genome⁴⁶. The absence of unique STS or fingerprint signatures over large genomic distances (~1 Mb) and the high degree of sequence similarity makes the distinction between paralogous sequence variation and allelic polymorphism problematic. Furthermore, the fact that such regions frequently harbour intron–exon structures of genuine unique sequence will complicate efforts to generate a genome-wide SNP map. The data indicate that a modest portion of the human genome may be relatively recalcitrant to genomic-based methods for SNP detection. Owing to their repetitive nature and their location in the genome, segmental

Table 16 Fraction of finished sequence in inter- and intrachromosomal duplications

Chromosome	Intrachromosomal (%)	Interchromosomal (%)	All (%)
1	1.4	0.5	1.9
2	0.1	0.6	0.7
3	0.3	1.1	1.1
4	0.0	1.0	1.0
5	0.6	0.3	0.9
6	0.8	0.4	1.1
7	3.4	1.3	4.1
8	0.3	0.1	0.3
9	0.8	2.9	3.7
10	2.1	0.8	2.9
11	1.2	2.1	2.3
12	1.5	0.3	1.8
13	0.0	0.5	0.5
14	0.6	0.4	1.0
15	3.0	6.9	6.9
16	4.5	2.0	5.8
17	1.6	0.3	1.8
18	0.0	0.7	0.7
19	3.6	0.3	3.8
20	0.2	0.3	0.5
21	1.4	1.6	3.0
22	6.1	2.6	7.5
X	1.8	3.2	5.0
Y	12.1	16.0	27.4
Un	0.0	0.5	0.5
Total	2.0	1.5	3.3

Excludes duplications with identities >99.5% to avoid artefactual duplication due to incomplete merger in the assembly process. Calculation was performed on the finished sequence available in September 2000 and reflects the duplications found within the total amount of finished sequence then. Note that there is some overlap between the interchromosomal and intrachromosomal sets.

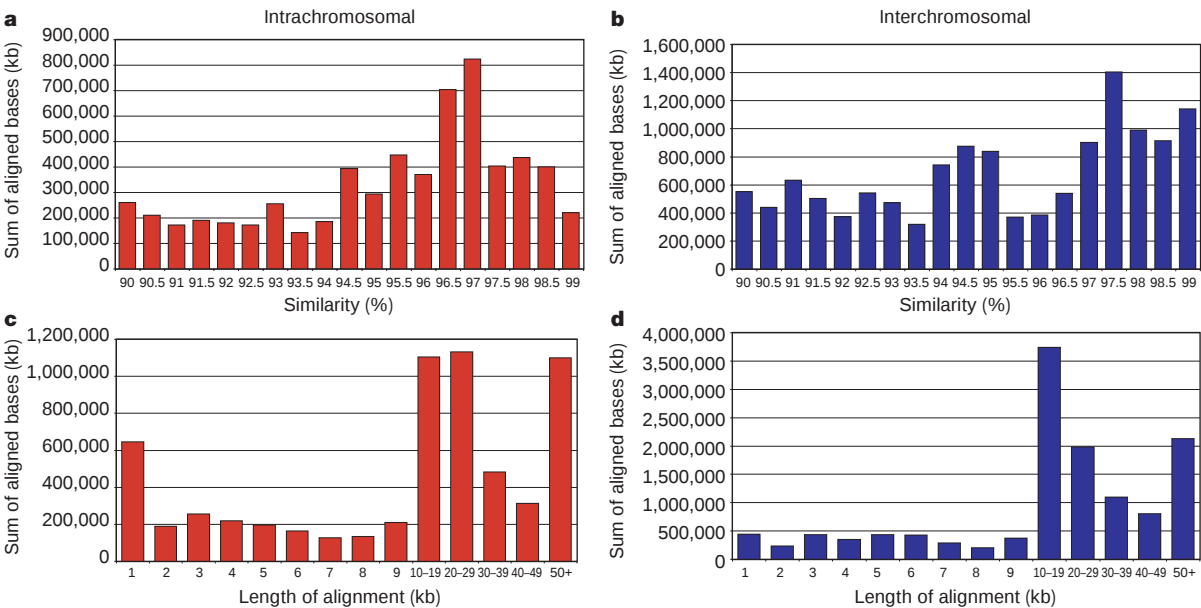


Figure 33 a–d, Sequence properties of segmental duplications. Distributions of length and per cent nucleotide identity for segmental duplications are shown as a function of the

number of aligned bp, for the subset of finished genome sequence. Intrachromosomal, red; interchromosomal, blue.

duplications may well be underestimated by the current analysis. An understanding of the biology, pathology and evolution of these duplications will require specialized efforts within these exceptional regions of the human genome. The presence and distribution of such segments may provide evolutionary fodder for processes of exon shuffling and a general increase in protein diversity associated with domain accretion. It will be important to consider both genome-wide duplication events and more restricted punctuated events of genome duplication as forces in the evolution of vertebrate genomes.

Table 17 Fraction of the draft genome sequence in inter- and intrachromosomal duplications

Chromosome	Intrachromosomal (%)	Interchromosomal (%)	All (%)
1	2.1	1.7	3.4
2	1.6	1.6	2.6
3	1.8	1.4	2.7
4	1.5	2.2	3.0
5	1.0	0.9	1.8
6	1.5	1.4	2.7
7	3.6	1.8	4.5
8	1.2	1.5	2.1
9	2.1	2.3	3.8
10	3.3	2.0	4.7
11	2.7	1.4	3.7
12	2.1	1.2	2.8
13	1.7	1.6	3.0
14	0.6	0.6	1.2
15	4.1	4.4	6.7
16	3.4	3.4	5.5
17	4.4	1.7	5.7
18	0.9	1.0	1.9
19	5.4	1.6	6.3
20	0.8	1.4	2.0
21	1.9	4.0	4.8
22	6.8	7.7	11.9
X	1.2	1.1	2.2
Y	10.9	13.1	20.8
NA	2.3	7.8	8.3
UL	11.6	20.8	22.2
Total	2.3	2.0	3.6

Excludes duplications with identities >98% to avoid artefactual duplication due to incomplete merger in the assembly process. Calculation was performed on an earlier version of the draft genome sequence based on data available in July 2000 and reflects the duplications found within the total amount of finished sequence then. Note that there is some overlap between the interchromosomal and intrachromosomal sets.

Gene content of the human genome

Genes (or at least their coding regions) comprise only a tiny fraction of human DNA, but they represent the major biological function of the genome and the main focus of interest by biologists. They are also the most challenging feature to identify in the human genome sequence.

The ultimate goal is to compile a complete list of all human genes and their encoded proteins, to serve as a ‘periodic table’ for biomedical research²⁴³. But this is a difficult task. In organisms with small genomes, it is straightforward to identify most genes by the presence of long ORFs. In contrast, human genes tend to have small exons (encoding an average of only 50 codons) separated by long introns (some exceeding 10 kb). This creates a signal-to-noise problem, with the result that computer programs for direct gene prediction have only limited accuracy. Instead, computational prediction of human genes must rely largely on the availability of cDNA sequences or on sequence conservation with genes and proteins from other organisms. This approach is adequate for strongly conserved genes (such as histones or ubiquitin), but may be less sensitive to rapidly evolving genes (including many crucial to speciation, sex determination and fertilization).

Here we describe our efforts to recognize both the RNA genes and protein-coding genes in the human genome. We also study the properties of the predicted human protein set, attempting to discern how the human proteome differs from those of invertebrates such as worm and fly.

Noncoding RNAs

Although biologists often speak of a tight coupling between ‘genes

Table 18 Cross-species comparison for large, highly homologous segmental duplications

	Percentage of genome (%)		
	Fly	Worm	Human (finished)*
> 1 kb	1.2	4.25	3.25
> 5 kb	0.37	1.50	2.86
> 10 kb	0.08	0.66	2.52

* This is an underestimate of the total amount of segmental duplication in the human genome because it only reflects duplication detectable with available finished sequence. The proportion of segmental duplications of > 1 kb is probably about 5% (see text).

and their encoded protein products', it is important to remember that thousands of human genes produce noncoding RNAs (ncRNAs) as their ultimate product²⁴⁴. There are several major classes of ncRNA. (1) Transfer RNAs (tRNAs) are the adapters that translate the triplet nucleic acid code of RNA into the amino-acid sequence of proteins; (2) ribosomal RNAs (rRNAs) are also central to the translational machinery, and recent X-ray crystallography results strongly indicate that peptide bond formation is catalysed by rRNA, not protein^{245,246}; (3) small nucleolar RNAs (snoRNAs) are required for rRNA processing and base modification in the nucleolus^{247,248}; and (4) small nuclear RNAs (snRNAs) are critical components of spliceosomes, the large ribonucleoprotein (RNP) complexes that splice introns out of pre-mRNAs in the nucleus. Humans have both a major, U2 snRNA-dependent spliceosome that splices most introns, and a minor, U12 snRNA-dependent spliceosome that splices a rare class of introns that often have AT/AC dinucleotides at the splice sites instead of the canonical GT/AG splice site consensus²⁴⁹.

Other ncRNAs include both RNAs of known biochemical function (such as telomerase RNA and the 7SL signal recognition particle RNA) and ncRNAs of enigmatic function (such as the large Xist transcript implicated in X dosage compensation²⁵⁰, or the small vault RNAs found in the bizarre vault ribonucleoprotein complex²⁵¹, which is three times the mass of the ribosome but has unknown function).

ncRNAs do not have translated ORFs, are often small and are not polyadenylated. Accordingly, novel ncRNAs cannot readily be found by computational gene-finding techniques (which search for features such as ORFs) or experimental sequencing of cDNA or EST libraries (most of which are prepared by reverse transcription using a primer complementary to a poly(A) tail). Even if the complete finished sequence of the human genome were available, discovering novel ncRNAs would still be challenging. We can, however, identify genomic sequences that are homologous to known ncRNA genes, using BLASTN or, in some cases, more specialized methods.

It is sometimes difficult to tell whether such homologous genes are orthologues, paralogues or closely related pseudogenes (because inactivating mutations are much less obvious than for protein-coding genes). For tRNA, there is sufficiently detailed information about the cloverleaf secondary structure to allow true genes and pseudogenes to be distinguished with high sensitivity. For many other ncRNAs, there is much less structural information and so we employ an operational criterion of high sequence similarity (> 95% sequence identity and > 95% full length) to distinguish true genes from pseudogenes. These assignments will eventually need to be reconciled with experimental data.

Transfer RNA genes. The classical experimental estimate of the number of human tRNA genes is 1,310 (ref. 252). In the draft genome sequence, we find only 497 human tRNA genes (Tables 19, 20). How do we account for this discrepancy? We believe that the original estimate is likely to have been inflated in two respects. First, it came from a hybridization experiment that probably counted closely related pseudogenes; by analysis of the draft genome sequence, there are in fact 324 tRNA-derived putative pseudogenes

(Table 20). Second, the earlier estimate assumed too high a value for the size of the human genome; repeating the calculation using the correct value yields an estimate of about 890 tRNA-related loci, which is in reasonable accord with our count of 821 tRNA genes and pseudogenes in the draft genome sequence.

The human tRNA gene set predicted from the draft genome sequence appears to include most of the known human tRNA species. The draft genome sequence contains 37 of 38 human tRNA species listed in a tRNA database²⁵³, allowing for up to one mismatch. This includes one copy of the known gene for a specialized selenocysteine tRNA, one of several components of a baroque translational mechanism that reads UGA as a selenocysteine codon in certain rare mRNAs that carry a specific *cis*-acting RNA regulatory site (a so-called SECIS element) in their 3' UTRs. The one tRNA gene in the database not found in the draft genome sequence is DE9990, a tRNA^{Glu} species, which differs in two positions from the most related tRNA gene in the human genome. Possible explanations are that the database version of this tRNA contains two errors, the gene is polymorphic or this is a genuine functional tRNA that is missing from the draft genome sequence. (The database also lists one additional tRNA gene (DS9994), but this is apparently a contaminant, most similar to bacterial tRNAs; the parent entry (Z13399) was withdrawn from the DNA database, but the tRNA entry has not yet been removed from the tRNA database.) Although the human set appears substantially complete by this test, the tRNA gene numbers in Table 19 should be considered tentative and used with caution. The human and fly (but not the worm) are known to be missing significant amounts of heterochromatic DNA, and additional tRNA genes could be located there.

With this caveat, the results indicate that the human has fewer tRNA genes than the worm, but more than the fly. This may seem surprising, but tRNA gene number in metazoans is thought to be related not to organismal complexity, but more to idiosyncrasies of the demand for tRNA abundance in certain tissues or stages of embryonic development. For example, the frog *Xenopus laevis*, which must load each oocyte with a remarkable 40 ng of tRNA, has thousands of tRNA genes²⁵⁴.

The degeneracy of the genetic code has allowed an inspired economy of tRNA anticodon usage. Although 61 sense codons need to be decoded, not all 61 different anticodons are present in tRNAs. Rather, tRNAs generally follow stereotyped and conserved wobble rules^{255–257}. Wobble reduces the number of required anticodons substantially, and provides a connection between the genetic code and the hybridization stability of modified and unmodified RNA bases. In eukaryotes, the rules proposed by Guthrie and Abelson²⁵⁶ predict that about 46 tRNA species will be sufficient to read the 61 sense codons (counting the initiator and elongator methionine tRNAs as two species). According to these rules, in the codon's third (wobble) position, U and C are generally decoded by a single tRNA species, whereas A and G are decoded by two separate tRNA species.

In 'two-codon boxes' of the genetic code (where codons ending with U/C encode a different amino acid from those ending with A/G), the U/C wobble position should be decoded by a G at position 34 in the tRNA anticodon. Thus, in the top left of Fig. 34, there is no tRNA with an AAA anticodon for Phe, but the GAA anticodon can recognize both UUU and UUC codons in the mRNA. In 'four-codon boxes' of the genetic code (where U, C, A and G in the wobble position all encode the same amino acid), the U/C wobble position is almost always decoded by I34 (inosine) in the tRNA, where the inosine is produced by post-transcriptional modification of an adenine (A). In the bottom left of Fig. 34, for example, the GUU and GUC codons of the four-codon Val box are decoded by a tRNA with an anticodon of AAC, which is no doubt modified to IAC. Presumably this pattern, which is strikingly conserved in eukaryotes, has to do with the fact that IA base pairs are also possible; thus

Table 19 Number of tRNA genes in various organisms

Organism	Number of canonical tRNAs	SeCys tRNA
Human	497	1
Worm	584	1
Fly	284	1
Yeast	273	0
<i>Methanococcus jannaschii</i>	36	1
<i>Escherichia coli</i>	86	1

Number of tRNA genes in each of six genome sequences, according to analysis by the computer program tRNAscan-SE. Canonical tRNAs read one of the standard 61 sense codons; this category excludes pseudogenes, undetermined anticodons, putative suppressors and selenocysteine tRNAs. Most organisms have a selenocysteine (SeCys) tRNA species, but some unicellular eukaryotes do not (such as the yeast *S. cerevisiae*).

the IAC anticodon for a Val tRNA could recognize GUU, GUC and even GUA codons. Were this same 134 to be utilized in two-codon boxes, however, misreading of the NNA codon would occur, resulting in translational havoc. Eukaryotic glycine tRNAs represent a conserved exception to this last rule; they use a GCC anticodon to decode GGU and GGC, rather than the expected ICC anticodon.

Satisfyingly, the human tRNA set follows these wobble rules almost perfectly (Fig. 34). Only three unexpected tRNA species are found: single genes for a tRNA^{Tyr}-AUA, tRNA^{Ile}-GAU, and tRNA^{Asn}-AUU. Perhaps these are pseudogenes, but they appear to be plausible tRNAs. We also checked the possibility of sequencing errors in their anticodons, but each of these three genes is in a region of high sequence accuracy, with PHRAP quality scores higher than 70 for every base in their anticodons.

As in all other organisms, human protein-coding genes show codon bias—preferential use of one synonymous codon over another²⁵⁸ (Fig. 34). In less complex organisms, such as yeast or bacteria, highly expressed genes show the strongest codon bias. Cytoplasmic abundance of tRNA species is correlated with both codon bias and overall amino-acid frequency (for example, tRNAs for preferred codons and for more common amino acids are more abundant). This is presumably driven by selective pressure for efficient or accurate translation²⁵⁹. In many organisms, tRNA abundance in turn appears to be roughly correlated with tRNA gene copy number, so tRNA gene copy number has been used as a proxy for tRNA abundance²⁶⁰. In vertebrates, however, codon bias is not so obviously correlated with gene expression level. Differing codon biases between human genes is more a function of their location in regions of different GC composition²⁶¹. In agreement with the literature, we see only a very rough correlation of human tRNA gene number with either amino-acid frequency or codon bias (Fig. 34). The most obvious outliers in these weak correlations are

the strongly preferred CUG leucine codon, with a mere six tRNA-Leu-CAG genes producing a tRNA to decode it, and the relatively rare cysteine UGU and UGC codons, with 30 tRNA genes to decode them.

The tRNA genes are dispersed throughout the human genome. However, this dispersal is nonrandom. tRNA genes have sometimes been seen in clusters at small scales^{262,263} but we can now see striking clustering on a genome-wide scale. More than 25% of the tRNA genes (140) are found in a region of only about 4 Mb on chromosome 6. This small region, only about 0.1% of the genome, contains an almost sufficient set of tRNA genes all by itself. The 140 tRNA genes contain a representative for 36 of the 49 anticodons found in the complete set; and of the 21 isoacceptor types, only tRNAs to decode Asn, Cys, Glu and selenocysteine are missing. Many of these tRNA genes, meanwhile, are clustered elsewhere; 18 of the 30 Cys tRNAs are found in a 0.5-Mb stretch of chromosome 7 and many of the Asn and Glu tRNA genes are loosely clustered on chromosome 1. More than half of the tRNA genes (280 out of 497) reside on either chromosome 1 or chromosome 6. Chromosomes 3, 4, 8, 9, 10, 12, 18, 20, 21 and X appear to have fewer than 10 tRNA genes each; and chromosomes 22 and Y have none at all (each has a single pseudogene).

Ribosomal RNA genes. The ribosome, the protein synthetic machine of the cell, is made up of two subunits and contains four rRNA species and many proteins. The large ribosomal subunit contains 28S and 5.8S rRNAs (collectively called 'large subunit' (LSU) rRNA) and also a 5S rRNA. The small ribosomal subunit contains 18S rRNA ('small subunit' (SSU) rRNA). The genes for LSU and SSU rRNA occur in the human genome as a 44-kb tandem repeat unit²⁶⁴. There are thought to be about 150–200 copies of this repeat unit arrayed on the short arms of acrocentric chromosomes 13, 14, 15, 21 and 22 (refs 254, 264). There are no true complete

Phe [171 UUU \ AAA 0 203 UUC \ GAA 14 Leu [73 UUA \ UAA 8 125 UUG \ CAA 6	Ser [147 UCU \ AGA 10 172 UCC \ GGA 0 118 UCA \ UGA 5 45 UCG \ CGA 4	Tyr [124 UAU \ AUA 1 158 UAC \ GUA 11 stop [0 UAA \ UUA 0 0 UAG \ CUA 0	Cys [99 UGU \ ACA 0 119 UGC \ GCA 30 stop [0 UGA \ UCA 0 Trp [122 UGG \ CCA 7
Leu [127 CUU \ AAG 13 187 CUC \ GAG 0 69 CUA \ UAG 2 392 CUG \ CAG 6	Pro [175 CCU \ AGG 11 197 CCC \ GGG 0 170 CCA \ UGG 10 69 CCG \ CGG 4	His [104 CAU \ AUG 0 147 CAC \ GUG 12 Gln [121 CAA \ UUG 11 343 CAG \ CUG 21	Arg [47 CGU \ ACG 9 107 CGC \ GCG 0 63 CGA \ UCG 7 115 CGG \ CCG 5
Ile [165 AUU \ AAU 13 218 AUC \ GAU 1 71 AUA \ UAU 5 Met [221 AUG \ CAU 17	Thr [131 ACU \ AGU 8 192 ACC \ GGU 0 150 ACA \ UGU 10 63 ACG \ CGU 7	Asn [174 AAU \ AUU 1 199 AAC \ GUU 33 Lys [248 AAA \ UUU 16 331 AAG \ CUU 22	Ser [121 AGU \ ACU 0 191 AGC \ GCU 7 Arg [113 AGA \ UCU 5 110 AGG \ CCU 4
Val [111 GUU \ AAC 20 146 GUC \ GAC 0 72 GUA \ UAC 5 288 GUG \ CAC 19	Ala [185 GCU \ AGC 25 282 GCC \ GGC 0 160 GCA \ UGC 10 74 GCG \ CGC 5	Asp [230 GAU \ AUC 0 262 GAC \ GUC 10 Glu [301 GAA \ UUC 14 404 GAG \ CUC 8	Gly [112 GGU \ ACC 0 230 GGC \ GCC 11 168 GGA \ UCC 5 160 GGG \ CCC 8

Figure 34 The human genetic code and associated tRNA genes. For each of the 64 codons, we show: the corresponding amino acid; the observed frequency of the codon per 10,000 codons; the codon; predicted wobble pairing to a tRNA anticodon (black lines); an unmodified tRNA anticodon sequence; and the number of tRNA genes found with this anticodon. For example, phenylalanine is encoded by UUU or UUC; UUC is seen more frequently, 203 to 171 occurrences per 10,000 total codons; both codons are expected to be decoded by a single tRNA anticodon type, GAA, using a G/U wobble; and there are 14

tRNA genes found with this anticodon. The modified anticodon sequence in the mature tRNA is not shown, even where post-transcriptional modifications can be confidently predicted (for example, when an A is used to decode a U/C third position, the A is almost certainly an inosine in the mature tRNA). The Figure also does not show the number of distinct tRNA species (such as distinct sequence families) for each anticodon; often there is more than one species for each anticodon.

copies of the rDNA tandem repeats in the draft genome sequence, owing to the deliberate bias in the initial phase of the sequencing effort against sequencing BAC clones whose restriction fragment fingerprints showed them to contain primarily tandemly repeated sequence. Sequence similarity analysis with the BLASTN computer program does, however, detect hundreds of rDNA-derived sequence fragments dispersed throughout the complete genome, including one 'full-length' copy of an individual 5.8S rRNA gene not associated with a true tandem repeat unit (Table 20).

The 5S rDNA genes also occur in tandem arrays, the largest of which is on chromosome 1 between 1q41.11 and 1q42.13, close to the telomere^{265,266}. There are 200–300 true 5S genes in these arrays^{265,267}. The number of 5S-related sequences in the genome, including numerous dispersed pseudogenes, is classically cited as 2,000 (refs 252, 254). The long tandem array on chromosome 1 is not yet present in the draft genome sequence because there are no *EcoRI* or *HindIII* sites present, and thus it was not cloned in the most heavily utilized BAC libraries (Table 1). We expect to recover it during the finishing stage. We do detect four individual copies of 5S rDNA by our search criteria ($\geq 95\%$ identity and $\geq 95\%$ full length). We also find many more distantly related dispersed sequences (520 at $P \leq 0.001$), which we interpret as probable pseudogenes (Table 20).

Small nucleolar RNA genes. Eukaryotic rRNA is extensively processed and modified in the nucleolus. Much of this activity is directed by numerous snoRNAs. These come in two families: C/D box snoRNAs (mostly involved in guiding site-specific 2'-O-ribose methylations of other RNAs) and H/ACA snoRNAs (mostly involved in guiding site-specific pseudouridylations)^{247,248}. We compiled a set of 97 known human snoRNA gene sequences; 84 of these (87%) have at least one copy in the draft genome sequence (Table 20), almost all as single-copy genes.

It is thought that all 2'-O-ribose methylations and pseudouridylations in eukaryotic rRNA are guided by snoRNAs. There are 105–107 methylations and around 95 pseudouridylations in human

rRNA²⁶⁸. Only about half of these have been tentatively assigned to known guide snoRNAs. There are also snoRNA-directed modifications on other stable RNAs, such as U6 (ref. 269), and the extent of this is just beginning to be explored. Sequence similarity has so far proven insufficient to recognize all snoRNA genes. We therefore expect that there are many unrecognized snoRNA genes that are not detected by BLAST queries.

Spliceosomal RNAs and other ncRNA genes. We also looked for copies of other known ncRNA genes. We found at least one copy of 21 (95%) of 22 known ncRNAs, including the spliceosomal snRNAs. There were multiple copies for several ncRNAs, as expected; for example, we find 44 dispersed genes for U6 snRNA, and 16 for U1 snRNA (Table 20).

For some of these RNA genes, homogeneous multigene families that occur in tandem arrays are again under-represented owing to the restriction enzymes used in constructing the BAC libraries and, in some instances, the decision to delay the sequencing of BAC clones with low complexity fingerprints indicative of tandemly repeated DNA. The U2 RNA genes are located at the RNU2 locus, a tandem array of 10–20 copies of nearly identical 6.1-kb units at 17q21–q22 (refs 270–272). Similarly, the U3 snoRNA genes (included in the aggregate count of C/D snoRNAs in Table 20) are clustered at the RNU3 locus at 17p11.2, not in a tandem array, but in a complex inverted repeat structure of about 5–10 copies per haploid genome²⁷³. The U1 RNA genes are clustered with about 30 copies at the RNU1 locus at 1p36.1, but this cluster is thought to be loose and irregularly organized; no two U1 genes have been cloned on the same cosmid²⁷¹. In the draft genome sequence, we see six copies of U2 RNA that meet our criteria for true genes, three of which appear to be in the expected position on chromosome 17. For U3, so far we see one true copy at the correct place on chromosome 17p11.2. For U1, we see 16 true genes, 6 of which are loosely clustered within 0.6 Mb at 1p36.1 and another 6 are elsewhere on chromosome 1. Again, these and other clusters will be a matter for the finishing process.

Table 20 Known non-coding RNA genes in the draft genome sequence

RNA gene*	Number expected†	Number found‡	Number of related genes§	Function
tRNA	1,310	497	324	Protein synthesis
SSU (18S) rRNA	150–200	0	40	Protein synthesis
5.8S rRNA	150–200	1	11	Protein synthesis
LSU (28S) rRNA	150–200	0	181	Protein synthesis
5S rRNA	200–300	4	520	Protein synthesis
U1	~30	16	134	Spliceosome component
U2	10–20	6	94	Spliceosome component
U4	??	4	87	Spliceosome component
U4atac	??	1	20	Component of minor (U11/U12) spliceosome
U5	??	1	31	Spliceosome component
U6	??	44	1,135	Spliceosome component
U6atac	??	4	32	Component of minor (U11/U12) spliceosome
U7	1	1	3	Histone mRNA 3' processing
U11	1	0	6	Component of minor (U11/U12) spliceosome
U12	1	1	0	Component of minor (U11/U12) spliceosome
SRP (7SL) RNA	4	3	773	Component of signal recognition particle (protein secretion)
RNAse P	1	1	2	tRNA 5' end processing
RNAse MRP	1	1	6	rRNA processing
Telomerase RNA	1	1	4	Template for addition of telomeres
hY1	1	1	353	Component of Ro RNP, function unknown
hY3	1	25	414	Component of Ro RNP, function unknown
hY4	1	3	115	Component of Ro RNP, function unknown
hY5 (4.5S RNA)	1	1	9	Component of Ro RNP, function unknown
Vault RNAs	3	3	1	Component of 13-MDa vault RNP, function unknown
7SK	1	1	330	Unknown
H19	1	1	2	Unknown
Xist	1	1	0	Initiation of X chromosome inactivation (dosage compensation)
Known C/D snoRNAs	81	69	558	Pre-rRNA processing or site-specific ribose methylation of rRNA
Known H/ACA snoRNAs	16	15	87	Pre-rRNA processing or site-specific pseudouridylation of rRNA

* Known ncRNA genes (or gene families, such as the C/D and H/ACA snoRNA families); reference sequences were extracted from GenBank and used to probe the draft genome sequence.

† Number of genes that were expected in the human genome, based on previous literature (note that earlier experimental techniques probably tend to overestimate copy number, by counting closely related pseudogenes).

‡ The copy number of 'true' full-length genes identified in the draft genome sequence.

§ The copy number of other significantly related copies (pseudogenes, fragments, paralogues) found. Except for the 497 true tRNA genes, all sequence similarities were identified by WashU BLASTN 2.0.MP (W. Gish, unpublished; <http://blast.wustl.edu>), with parameters '-kap wordmask = seg B = 50000 W = 8' and the default +5/-4 DNA scoring matrix. True genes were operationally defined as BLAST hits with $\geq 95\%$ identity over $\geq 95\%$ of the length of the query. Related sequences were operationally defined as all other BLAST hits with P -values ≤ 0.001 .

Table 21 Characteristics of human genes

	Median	Mean	Sample (size)
Internal exon	122 bp	145 bp	RefSeq alignments to draft genome sequence, with confirmed intron boundaries (43,317 exons)
Exon number	7	8.8	RefSeq alignments to finished sequence (3,501 genes)
Introns	1,023 bp	3,365 bp	RefSeq alignments to finished sequence (27,238 introns)
3' UTR	400 bp	770 bp	Confirmed by mRNA or EST on chromosome 22 (689)
5' UTR	240 bp	300 bp	Confirmed by mRNA or EST on chromosome 22 (463)
Coding sequence (CDS)	1,100 bp	1,340 bp	Selected RefSeq entries (1,804)
Genomic extent	367 aa	447 aa	
	14 kb	27 kb	Selected RefSeq entries (1,804)

Median and mean values for a number of properties of human protein-coding genes. The 1,804 selected RefSeq entries were those that could be unambiguously aligned to finished sequence over their entire length.

Our observations also confirm the striking proliferation of ncRNA-derived pseudogenes (Table 20). There are hundreds or thousands of sequences in the draft genome sequence related to some of the ncRNA genes. The most prolific pseudogene counts generally come from RNA genes transcribed by RNA polymerase III promoters, including U6, the hY RNAs and SRP-RNA. These ncRNA pseudogenes presumably arise through reverse transcription. The frequency of such events gives insight into how ncRNA genes can evolve into SINE retroposons, such as the tRNA-derived SINEs found in many vertebrates and the SRP-RNA-derived Alu elements found in humans.

Protein-coding genes

Identifying the protein-coding genes in the human genome is one of the most important applications of the sequence data, but also one of the most difficult challenges. We describe below our efforts to create an initial human gene and protein index.

Exploring properties of known genes. Before attempting to identify new genes, we explored what could be learned by aligning the cDNA sequences of known genes to the draft genome sequence. Genomic alignments allow one to study exon–intron structure and local GC content, and are valuable for biomedical studies because they connect genes with the genetic and cytogenetic map, link them with regulatory sequences and facilitate the development of polymerase chain reaction (PCR) primers to amplify exons. Until now, genomic alignment was available for only about a quarter of known genes.

The ‘known’ genes studied were those in the RefSeq database¹¹⁰, a manually curated collection designed to contain nonredundant representatives of most full-length human mRNA sequences in GenBank (RefSeq intentionally contains some alternative splice forms of the same genes). The version of RefSeq used contained 10,272 mRNAs.

The RefSeq genes were aligned with the draft genome sequence, using both the Spidey (S. Wheelan, personal communication) and Asembly (D. Thierry-Mieg and J. Thierry-Mieg, unpublished; <http://www.acedb.org>) computer programs. Because this sequence is incomplete and contains errors, not all genes could be fully aligned and some may have been incorrectly aligned. More than 92% of the RefSeq entries could be aligned at high stringency over at least part of their length, and 85% could be aligned over more than half of their length. Some genes (16%) had high stringency alignments to more than one location in the draft genome sequence owing, for example, to paralogues or pseudogenes. In such cases, we considered only the best match. In a few of these cases, the assignment may not be correct because the true matching region has not yet been sequenced. Three per cent of entries appeared to be alternative splice products of the same gene, on the basis of their alignment to the same location in the draft genome sequence. In all, we obtained at least partial genomic alignments for 9,212 distinct known genes and essentially complete alignment for 5,364 of them.

Previous efforts to study human gene structure^{116,274,275} have been hampered by limited sample sizes and strong biases in favour of compact genes. Table 21 gives the mean and median values of some basic characteristics of gene structures. Some of the values may be

underestimates. In particular, the UTRs given in the RefSeq database are likely to be incomplete; they are considerably shorter, for example, than those derived from careful reconstructions on chromosome 22. Intron sizes were measured only for genes in finished genomic sequence, to mitigate the bias arising from the fact that

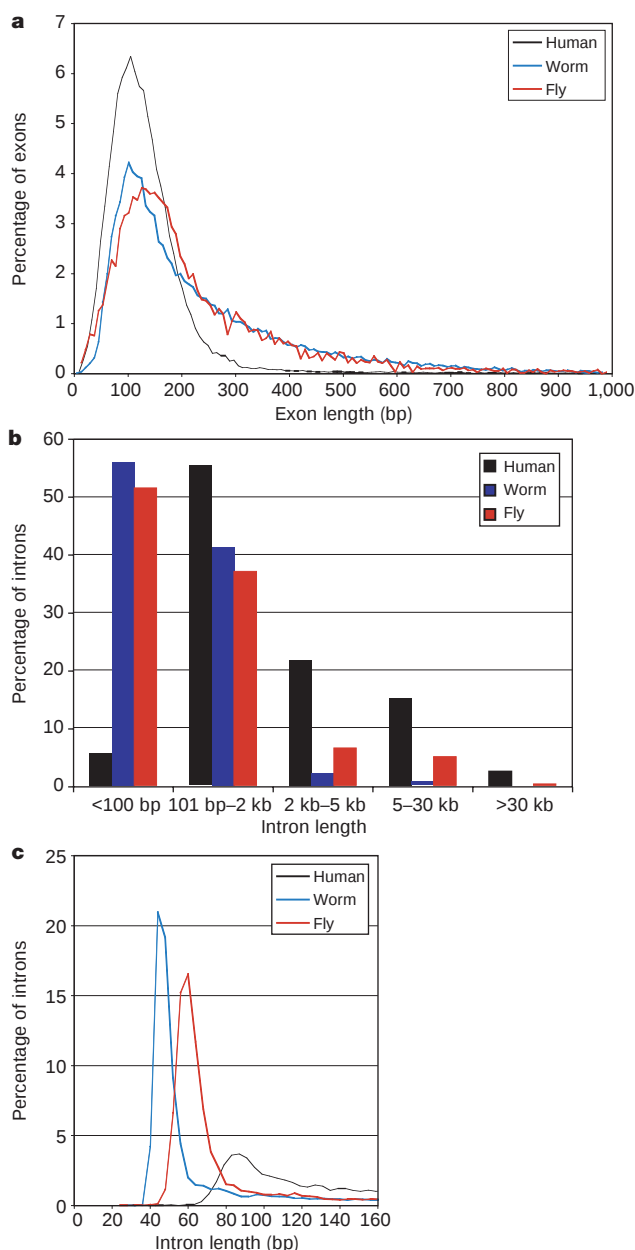


Figure 35 Size distributions of exons, introns and short introns, in sequenced genomes. **a**, Exons; **b**, introns; **c**, short introns (enlarged from **b**). Confirmed exons and introns for the human were taken from RefSeq alignments and for worm and fly from Asembly alignments of ESTs (J. and D. Thierry-Mieg and, for worm, Y. Kohara, unpublished).

long introns are more likely than short introns to be interrupted by gaps in the draft genome sequence. Nonetheless, there may be some residual bias against long genes and long introns.

There is considerable variation in overall gene size and intron size, with both distributions having very long tails. Many genes are over 100 kb long, the largest known example being the dystrophin gene (DMD) at 2.4 Mb. The variation in the size distribution of coding sequences and exons is less extreme, although there are still some remarkable outliers. The titin gene²⁷⁶ has the longest currently known coding sequence at 80,780 bp; it also has the largest number of exons (178) and longest single exon (17,106 bp).

It is instructive to compare the properties of human genes with those from worm and fly. For all three organisms, the typical length of a coding sequence is similar (1,311 bp for worm, 1,497 bp for fly and 1,340 bp for human), and most internal exons fall within a common peak between 50 and 200 bp (Fig. 35a). However, the worm and fly exon distributions have a fatter tail, resulting in a larger mean size for internal exons (218 bp for worm versus 145 bp for human). The conservation of preferred exon size across all three species supports suggestions of a conserved exon-based component of the splicing machinery²⁷⁷. Intriguingly, the few extremely short human exons show an unusual base composition. In 42 detected

human exons of less than 19 bp, the nucleotide frequencies of A, G, T and C are 39, 33, 15 and 12%, respectively, showing a strong purine bias. Purine-rich sequences may enhance splicing^{278,279}, and it is possible that such sequences are required or strongly selected for to ensure correct splicing of very short exons. Previous studies have shown that short exons require intronic, but not exonic, splicing enhancers²⁸⁰.

In contrast to the exons, the intron size distributions differ substantially among the three species (Fig. 35b, c). The worm and fly each have a reasonably tight distribution, with most introns near the preferred minimum intron length (47 bp for worm, 59 bp for fly) and an extended tail (overall average length of 267 bp for worm and 487 bp for fly). Intron size is much more variable in humans, with a peak at 87 bp but a very long tail resulting in a mean of more than 3,300 bp. The variation in intron size results in great variation in gene size.

The variation in gene size and intron size can partly be explained by the fact that GC-rich regions tend to be gene-dense with many compact genes, whereas AT-rich regions tend to be gene-poor with many sprawling genes containing large introns. The correlation of gene density with GC content is shown in Fig. 36a, b; the relative density increases more than tenfold as GC content increases from

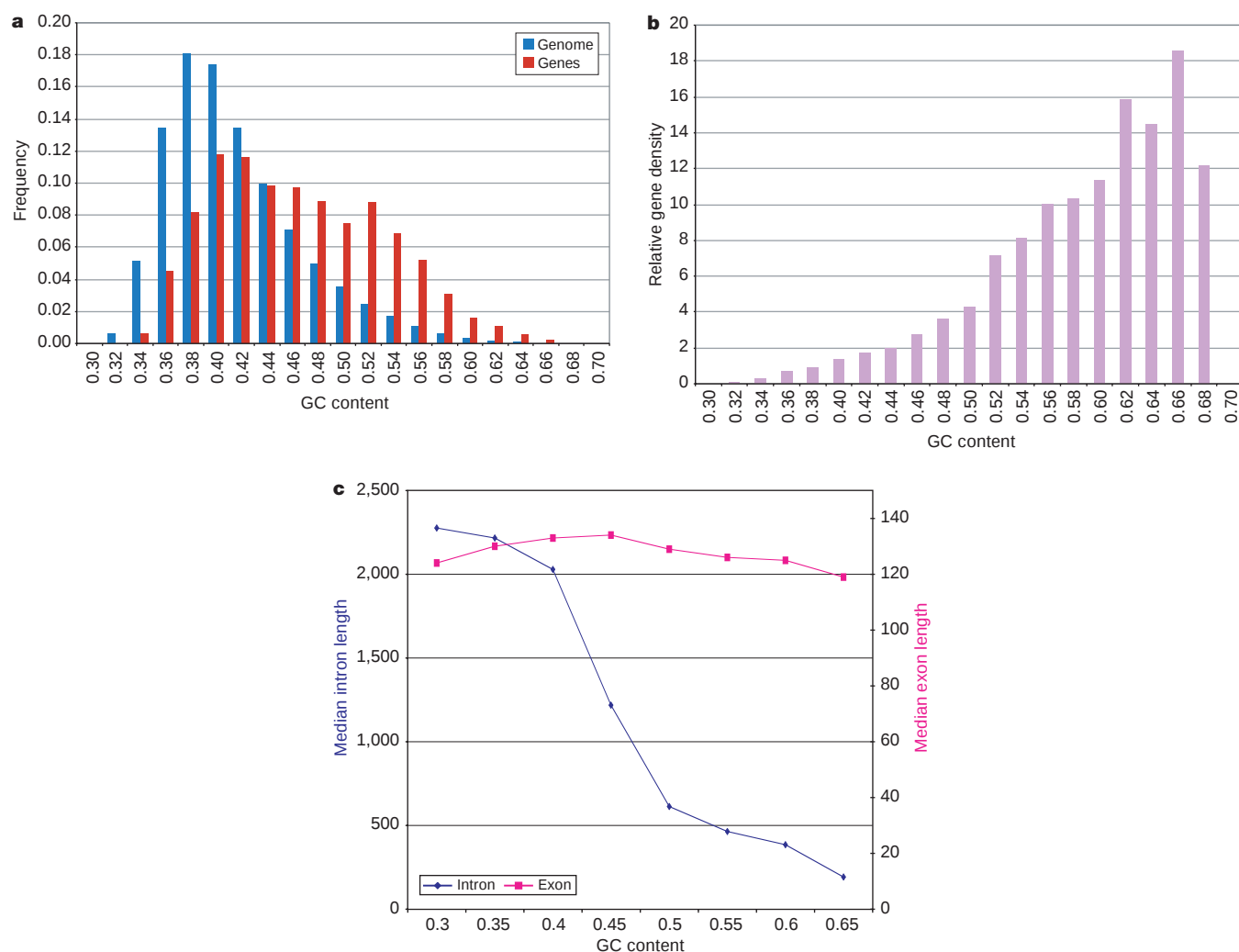


Figure 36 GC content. **a**, Distribution of GC content in genes and in the genome. For 9,315 known genes mapped to the draft genome sequence, the local GC content was calculated in a window covering either the whole alignment or 20,000 bp centred around the midpoint of the alignment, whichever was larger. Ns in the sequence were not counted. GC content for the genome was calculated for adjacent nonoverlapping 20,000-bp windows across the sequence. Both the gene and genome distributions have been

normalized to sum to one. **b**, Gene density as a function of GC content, obtained by taking the ratio of the data in **a**. Values are less accurate at higher GC levels because the denominator is small. **c**, Dependence of mean exon and intron lengths on GC content. For exons and introns, the local GC content was derived from alignments to finished sequence only, and were calculated from windows covering the feature or 10,000 bp centred on the feature, whichever was larger.

30% to 50%. The correlation appears to be due primarily to intron size, which drops markedly with increasing GC content (Fig. 36c). In contrast, coding properties such as exon length (Fig. 36c) or exon number (data not shown) vary little. Intergenic distance is also probably lower in high-GC areas, although this is hard to prove directly until all genes have been identified.

The large number of confirmed human introns allows us to analyse variant splice sites, confirming and extending recent reports²⁸¹. Intron positions were confirmed by applying a stringent criterion that EST or mRNA sequence show an exact match of 8 bp in the flanking exonic sequence on each side. Of 53,295 confirmed introns, 98.12% use the canonical dinucleotides GT at the 5' splice site and AG at the 3' site (GT-AG pattern). Another 0.76% use the related GC-AG. About 0.10% use AT-AC, which is a rare alternative pattern primarily recognized by the variant U12 splicing machinery²⁸². The remaining 1% belong to 177 types, some of which undoubtedly reflect sequencing or alignment errors.

Finally, we looked at alternative splicing of human genes. Alternative splicing can allow many proteins to be produced from a single gene and can be used for complex gene regulation. It appears to be prevalent in humans, with lower estimates of about 35% of human genes being subject to alternative splicing²⁸³⁻²⁸⁵. These studies may have underestimated the prevalence of alternative splicing, because they examined only EST alignments covering only a portion of a gene.

To investigate the prevalence of alternative splicing, we analysed reconstructed mRNA transcripts covering the entire coding regions of genes on chromosome 22 (omitting small genes with coding regions of less than 240 bp). Potential transcripts identified by alignments of ESTs and cDNAs to genomic sequence were verified by human inspection. We found 642 transcripts, covering 245 genes (average of 2.6 distinct transcripts per gene). Two or more alternatively spliced transcripts were found for 145 (59%) of these genes. A similar analysis for the gene-rich chromosome 19 gave 1,859 transcripts, corresponding to 544 genes (average 3.2 distinct transcripts per gene). Because we are sampling only a subset of all transcripts, the true extent of alternative splicing is likely to be greater. These figures are considerably higher than those for worm, in which analysis reveals alternative splicing for 22% of genes for which ESTs have been found, with an average of 1.34 (12,816/9,516) splice variants per gene. (The apparently higher extent of alternative splicing seen in human than in worm was not an artefact resulting from much deeper coverage of human genes by ESTs and mRNAs. Although there are many times more ESTs available for human than worm, these ESTs tend to have shorter average length (because many were the product of early sequencing efforts) and many match no human genes. We calculated the actual coverage per bp used in the analysis of the human and worm genes; the coverage is only modestly higher (about 50%) for the human, with a strong bias towards 3' UTRs which tend to show much less alternative splicing. We also repeated the analysis using equal coverage for the two organisms and confirmed that higher levels of alternative splicing were still seen in human.)

Seventy per cent of alternative splice forms found in the genes on chromosomes 19 and 22 affect the coding sequence, rather than merely changing the 3' or 5' UTR. (This estimate may be affected by the incomplete representation of UTRs in the RefSeq database and in the transcripts studied.) Alternative splicing of the terminal exon was seen for 20% of 6,105 mRNAs that were aligned to the draft genome sequence and correspond to confirmed 3' EST clusters. In addition to alternative splicing, we found evidence of the terminal exon employing alternative polyadenylation sites (separated by > 100 bp) in 24% of cases.

Towards a complete index of human genes. We next focused on creating an initial index of human genes and proteins. This index is quite incomplete, owing to the difficulty of gene identification in human DNA and the imperfect state of the draft genome sequence.

Nonetheless, it is valuable for experimental studies and provides important insights into the nature of human genes and proteins.

The challenge of identifying genes from genomic sequence varies greatly among organisms. Gene identification is almost trivial in bacteria and yeast, because the absence of introns in bacteria and their paucity in yeast means that most genes can be readily recognized by *ab initio* analysis as unusually long ORFs. It is not as simple, but still relatively straightforward, to identify genes in animals with small genomes and small introns, such as worm and fly. A major factor is the high signal-to-noise ratio—coding sequences comprise a large proportion of the genome and a large proportion of each gene (about 50% for worm and fly), and exons are relatively large.

Gene identification is more difficult in human DNA. The signal-to-noise ratio is lower: coding sequences comprise only a few per cent of the genome and an average of about 5% of each gene; internal exons are smaller than in worms; and genes appear to have more alternative splicing. The challenge is underscored by the work on human chromosomes 21 and 22. Even with the availability of finished sequence and intensive experimental work, the gene content remains uncertain, with upper and lower estimates differing by as much as 30%. The initial report of the finished sequence of chromosome 22 (ref. 94) identified 247 previously known genes, 298 predicted genes confirmed by sequence homology or ESTs and 325 *ab initio* predictions without additional support. Many of the confirmed predictions represented partial genes. In the past year, 440 additional exons (10%) have been added to existing gene annotations by the chromosome 22 annotation group, although the number of confirmed genes has increased by only 17 and some previously identified gene predictions have been merged²⁸⁶.

Before discussing the gene predictions for the human genome, it is useful to consider background issues, including previous estimates of the number of human genes, lessons learned from worms and flies and the representativeness of currently 'known' human genes.

Previous estimates of human gene number. Although direct enumeration of human genes is only now becoming possible with the advent of the draft genome sequence, there have been many attempts in the past quarter of a century to estimate the number of genes indirectly. Early estimates based on reassociation kinetics estimated the mRNA complexity of typical vertebrate tissues to be 10,000–20,000, and were extrapolated to suggest around 40,000 for the entire genome²⁸⁷. In the mid-1980s, Gilbert suggested that there might be about 100,000 genes, based on the approximate ratio of the size of a typical gene ($\sim 3 \times 10^4$ bp) to the size of the genome (3×10^9 bp). Although this was intended only as a back-of-the-envelope estimate, the pleasing roundness of the figure seems to have led to it being widely quoted and adopted in many textbooks. (W. Gilbert, personal communication; ref. 288). An estimate of 70,000–80,000 genes was made by extrapolating from the number of CpG islands and the frequency of their association with known genes¹²⁹.

As human sequence information has accumulated, it has been possible to derive estimates on the basis of sampling techniques²⁸⁹. Such studies have sought to extrapolate from various types of data, including ESTs, mRNAs from known genes, cross-species genome comparisons and analysis of finished chromosomes. Estimates based on ESTs²⁹⁰ have varied widely, from 35,000 (ref. 130) to 120,000 genes²⁹¹. Some of the discrepancy lies in differing estimates of the amount of contaminating genomic sequence in the EST collection and the extent to which multiple distinct ESTs correspond to a single gene. The most rigorous analyses¹³⁰ exclude as spurious any ESTs that appear only once in the data set and carefully calibrate sensitivity and specificity. Such calculations consistently produce low estimates, in the region of 35,000.

Comparison of whole-genome shotgun sequence from the pufferfish *T. nigroviridis* with the human genome²⁹² can be used to estimate the density of exons (detected as conserved sequences

between fish and human). These analyses also suggest around 30,000 human genes.

Extrapolations have also been made from the gene counts for chromosomes 21 and 22 (refs 93, 94), adjusted for differences in gene densities on these chromosomes, as inferred from EST mapping. These estimates are between 30,500 and 35,500, depending on the precise assumptions used²⁸⁶.

Insights from invertebrates. The worm and fly genomes contain a large proportion of novel genes (around 50% of worm genes and 30% of fly genes), in the sense of showing no significant similarity to organisms outside their phylum^{293–295}. Such genes may have been present in the original eukaryotic ancestor, but were subsequently lost from the lineages of the other eukaryotes for which sequence is available; they may be rapidly diverging genes, so that it is difficult to recognize homologues solely on the basis of sequence; they may represent true innovations developed within the lineage; or they may represent acquisitions by horizontal transfer. Whatever their origin, these genes tend to have different biological properties from highly conserved genes. In particular, they tend to have low expression levels as assayed both by direct studies and by a paucity of corresponding ESTs, and are less likely to produce a visible phenotype in loss-of-function genetic experiments^{294,296}.

Gene prediction. Current gene prediction methods employ combinations of three basic approaches: direct evidence of transcription provided by ESTs or mRNAs^{297–299}; indirect evidence based on sequence similarity to previously identified genes and proteins^{300,301}; and *ab initio* recognition of groups of exons on the basis of hidden Markov models (HMMs) that combine statistical information about splice sites, coding bias and exon and intron lengths (for example, Genscan²⁷⁵, Genie^{302,303} and FGENES³⁰⁴).

The first approach relies on direct experimental data, but is subject to artefacts arising from contaminating ESTs derived from unspliced mRNAs, genomic DNA contamination and nongenic transcription (for example, from the promoter of a transposable element). The first two problems can be mitigated by comparing transcripts with the genomic sequence and using only those that show clear evidence of splicing. This solution, however, tends to discard evidence from genes with long terminal exons or single exons. The second approach tends correctly to identify gene-derived sequences, although some of these may be pseudogenes. However, it obviously cannot identify truly novel genes that have no sequence similarity to known genes. The third approach would suffice alone if one could accurately define the features used by cells for gene recognition, but our current understanding is insufficient to do so. The sensitivity and specificity of *ab initio* predictions are greatly affected by the signal-to-noise ratio. Such methods are more accurate in the fly and worm than in human. In fly, *ab initio* methods can correctly predict around 90% of individual exons and can correctly predict all coding exons of a gene in about 40% of cases³⁰³. For human, the comparable figures are only about 70% and 20%, respectively^{94,305}. These estimates may be optimistic, owing to the design of the tests used.

In any collection of gene predictions, we can expect to see various errors. Some gene predictions may represent partial genes, because of inability to detect some portions of a gene (incomplete sensitivity) or to connect all the components of a gene (fragmentation); some may be gene fusions; and others may be spurious predictions (incomplete specificity) resulting from chance matches or pseudo-genes.

Creating an initial gene index. We set out to create an initial integrated gene index (IGI) and an associated integrated protein index (IPI) for the human genome. We describe the results obtained from a version of the draft genome sequence based on the sequence data available in July 2000, to allow time for detailed analysis of the gene and protein content. The additional sequence data that has since become available will affect the results quantitatively, but are unlikely to change the conclusions qualitatively.

We began with predictions produced by the Ensembl system³⁰⁶. Ensembl starts with *ab initio* predictions produced by Genscan²⁷⁵ and then attempts to confirm them by virtue of similarity to proteins, mRNAs, ESTs and protein motifs (contained in the Pfam database³⁰⁷) from any organism. In particular, it confirms introns if they are bridged by matches and exons if they are flanked by confirmed introns. It then attempts to extend protein matches using the GeneWise computer program³⁰⁸. Because it requires confirmatory evidence to support each gene component, it frequently produces partial gene predictions. In addition, when there is evidence of alternative splicing, it reports multiple overlapping transcripts. In total, Ensembl produced 35,500 gene predictions with 44,860 transcripts.

To reduce fragmentation, we next merged Ensembl-based gene predictions with overlapping gene predictions from another program, Genie³⁰². Genie starts with mRNA or EST matches and employs an HMM to extend these matches by using *ab initio* statistical approaches. To avoid fragmentation, it attempts to link information from 5' and 3' ESTs from the same cDNA clone and thereby to produce a complete coding sequence from an initial ATG to a stop codon. As a result, it may generate complete genes more accurately than Ensembl in cases where there is extensive EST support. (Genie also generates potential alternative transcripts, but we used only the longest transcript in each group.) We merged 15,437 Ensembl predictions into 9,526 clusters, and the longest transcript in each cluster (from either Genie or Ensembl) was taken as the representative.

Next, we merged these results with known genes contained in the RefSeq (version of 29 September 2000), SWISSPROT (release 39.6 of 30 August 2000) and TrEMBL databases (TrEMBL release 14.17 of 1 October 2000, TrEMBL_new of 1 October 2000). Incorporating these sequences gave rise to overlapping sequences because of alternative splice forms and partial sequences. To construct a nonredundant set, we selected the longest sequence from each overlapping set by using direct protein comparison and by mapping the gene predictions back onto the genome to construct the overlapping sets. This may occasionally remove some close paralogues in the event that the correct genomic location has not yet been sequenced, but this number is expected to be small.

Finally, we searched the set to eliminate any genes derived from contaminating bacterial sequences, recognized by virtue of near identity to known bacterial plasmids, transposons and chromosomal genes. Although most instances of such contamination had been removed in the assembly process, a few cases had slipped through and were removed at this stage.

The process resulted in version 1 of the IGI (IGI.1). The composition of the corresponding IPI.1 protein set, obtained by translating IGI.1, is given in Table 22. There are 31,778 protein predictions, with 14,882 from known genes, 4,057 predictions from Ensembl merged with Genie and 12,839 predictions from Ensembl alone. The average lengths are 469 amino acids for the known proteins, 443 amino acids for protein predictions from the Ensembl–Genie merge, and 187 amino acids for those from Ensembl alone. (The smaller average size for the predictions from Ensembl alone reflects its tendency to predict partial genes where there is supporting evidence for only part of the gene; the remainder of the gene will often not be predicted at all, rather than included as part of another prediction. Accordingly, the smaller size cannot be used to estimate the rate of fragmentation in such predictions.)

The set corresponds to fewer than 31,000 actual genes, because some genes are fragmented into more than one partial prediction and some predictions may be spurious or correspond to pseudo-genes. As discussed below, our best estimate is that IGI.1 includes about 24,500 true genes.

Evaluation of IGI/IPI. We used several approaches to evaluate the sensitivity, specificity and fragmentation of the IGI/IPI set.

Comparison with 'new' known genes. One approach was to examine

Table 22 Properties of the IGI/IPI human protein set

Source	Number	Average length (amino acids)	Matches to nonhuman proteins	Matches to RIKEN mouse cDNA set	Matches to RIKEN mouse cDNA set but not to nonhuman proteins
RefSeq/SwissProt/TrEMBL	14,882	469	12,708 (85%)	11,599 (78%)	776 (36%)
Ensembl-Genie	4,057	443	2,989 (74%)	3,016 (74%)	498 (47%)
Ensembl	12,839	187	81,126 (63%)	7,372 (57%)	1,449 (31%)
Total	31,778	352	23,813 (75%)	219,873 (69%)	2,723 (34%)

The matches to nonhuman proteins were obtained by using Smith-Waterman sequence alignment with an E -value threshold of 10^{-3} and the matches to the RIKEN mouse cDNAs by using TBLASTN with an E -value threshold of 10^{-4} . The last column shows that a significant number of the IGI members that do not have nonhuman protein matches do match sequences in the RIKEN mouse cDNA set, suggesting that both the IGI and the RIKEN sets contain a significant number of novel proteins.

newly discovered genes arising from independent work that were not used in our gene prediction effort. We identified 31 such genes: 22 recent entries to RefSeq and 9 from the Sanger Centre's gene identification program on chromosome X. Of these, 28 were contained in the draft genome sequence and 19 were represented in the IGI/IPI. This suggests that the gene prediction process has a sensitivity of about 68% (19/28) for the detection of novel genes in the draft genome sequence and that the current IGI contains about 61% (19/31) of novel genes in the human genome. On average, 79% of each gene was detected. The extent of fragmentation could also be estimated: 14 of the genes corresponded to a single prediction in the IGI/IPI, three genes corresponded to two predictions, one gene to three predictions and one gene to four predictions. This corresponds to a fragmentation rate of about 1.4 gene predictions per true gene.

Comparison with RIKEN mouse cDNAs. In a less direct but larger-scale approach, we compared the IGI gene set to a set of mouse cDNAs sequenced by the Genome Exploration Group of the RIKEN Genomic Sciences Center³⁰⁹. This set of 15,294 cDNAs, subjected to full-insert sequencing, was enriched for novel genes by selecting cDNAs with novel 3' ends from a collection of nearly one million ESTs from diverse tissues and developmental timepoints. We determined the proportion of the RIKEN cDNAs that showed sequence similarity to the draft genome sequence and the proportion that showed sequence similarity to the IGI/IPI. Around 81% of the genes in the RIKEN mouse set showed sequence similarity to the human genome sequence, whereas 69% showed sequence similarity to the IGI/IPI. This suggests a sensitivity of 85% (69/81). This is higher than the sensitivity estimate above, perhaps because some of the matches may be due to paralogues rather than orthologues. It is consistent with the IGI/IPI representing a substantial fraction of the human proteome.

Conversely, 69% (22,013/31,898) of the IGI matches the RIKEN cDNA set. Table 22 shows the breakdown of these matches among the different components of the IGI. This is lower than the proportion of matches among known proteins, although this is expected because known proteins tend to be more highly conserved (see above) and because the predictions are on average shorter than known proteins. Table 22 also shows the numbers of matches to the RIKEN cDNAs among IGI members that do not match known proteins. The results indicate that both the IGI and the RIKEN set contain a significant number of genes that are novel in the sense of not having known protein homologues.

Comparison with genes on chromosome 22. We also compared the IGI/IPI with the gene annotations on chromosome 22, to assess the proportion of gene predictions corresponding to pseudogenes and to estimate the rate of overprediction. We compared 477 IGI gene predictions to 539 confirmed genes and 133 pseudogenes on chromosome 22 (with the immunoglobulin lambda locus excluded owing to its highly atypical gene structure). Of these, 43 hit 36 annotated pseudogenes. This suggests that 9% of the IGI predictions may correspond to pseudogenes and also suggests a fragmentation rate of 1.2 gene predictions per gene. Of the remaining hits, 63 did not overlap with any current annotations. This would suggest a rate of spurious predictions of about 13% (63/477), although the true rate is likely to be much lower because many of these may correspond to

unannotated portions of existing gene predictions or to currently unannotated genes (of which there are estimated to be about 100 on this chromosome⁹⁴).

Chromosomal distribution. Finally, we examined the chromosomal distribution of the IGI gene set. The average density of gene predictions is 11.1 per Mb across the genome, with the extremes being chromosome 19 at 26.8 per Mb and chromosome Y at 6.4 per Mb. It is likely that a significant number of the predictions on chromosome Y are pseudogenes (this chromosome is known to be rich in pseudogenes) and thus that the density for chromosome Y is an overestimate. The density of both genes and Alus on chromosome 19 is much higher than expected, even accounting for the high GC content of the chromosome; this supports the idea that Alu density is more closely correlated with gene density than with GC content itself.

Summary. We are clearly still some way from having a complete set of human genes. The current IGI contains significant numbers of partial genes, fragmented and fused genes, pseudogenes and spurious predictions, and it also lacks significant numbers of true genes. This reflects the current state of gene prediction methods in vertebrates even in finished sequence, as well as the additional challenges related to the current state of the draft genome sequence. Nonetheless, the gene predictions provide a valuable starting point for a wide range of biological studies and will be rapidly refined in the coming year.

The analysis above allows us to estimate the number of distinct genes in the IGI, as well as the number of genes in the human genome. The IGI set contains about 15,000 known genes and about 17,000 gene predictions. Assuming that the gene predictions are subject to a rate of overprediction (spurious predictions and pseudogenes) of 20% and a rate of fragmentation of 1.4, the IGI would be estimated to contain about 24,500 actual human genes. Assuming that the gene predictions contain about 60% of previously unknown human genes, the total number of genes in the human genome would be estimated to be about 31,000. This is consistent with most recent estimates based on sampling, which suggest a gene number of 30,000–35,000. If there are 30,000–35,000 genes, with an average coding length of about 1,400 bp and average genomic extent of about 30 kb, then about 1.5% of the human genome would consist of coding sequence and one-third of the genome would be transcribed in genes.

The IGI/IPI was constructed primarily on the basis of gene predictions from Ensembl. However, we also generated an expanded set (IGI+) by including additional predictions from two other gene prediction programs, Genie and GenomeScan (C. Burge, personal communication). These predictions were not included in the core IGI set, because of the concern that each additional set will provide diminishing returns in identifying true genes while contributing its own false positives (increased sensitivity at the expense of specificity). Genie produced an additional 2,837 gene predictions not overlapping the IGI, and GenomeScan produced 6,534 such gene predictions. If all of these gene predictions were included in the IGI, the number of the 31 new 'known' genes (see above) contained in the IGI would rise from 19 to 24. This would amount to an increase of about 26% in sensitivity, at the expense of increasing the number of predicted genes (excluding knowns) by 55%. Allowing a higher

overprediction rate of 30% for gene predictions in this expanded set, the analysis above suggests that IGI+ set contains about 28,000 true genes and yields an estimate of about 32,000 human genes. We are investigating ways to filter the expanded set, to produce an IGI with the advantage of the increased sensitivity resulting from combining multiple gene prediction programs without the corresponding loss of specificity. Meanwhile, the IGI+ set can be used by researchers searching for genes that cannot be found in the IGI.

Some classes of genes may have been missed by all of the gene-finding methods. Genes could be missed if they are expressed at low levels or in rare tissues (being absent or very under-represented in EST and mRNA databases) and have sequences that evolve rapidly (being hard to detect by protein homology and genome comparison). Both the worm and fly gene sets contain a substantial number of such genes^{293,294}. Single-exon genes encoding small proteins may also have been missed, because EST evidence that supports them cannot be distinguished from genomic contamination in the EST dataset and because homology may be hard to detect for small proteins³¹⁰.

The human thus appears to have only about twice as many genes as worm or fly. However, human genes differ in important respects from those in worm and fly. They are spread out over much larger regions of genomic DNA, and they are used to construct more alternative transcripts. This may result in perhaps five times as many primary protein products in the human as in the worm or fly.

The predicted gene and protein sets described here are clearly far from final. Nonetheless, they provide a valuable starting point for experimental and computational research. The predictions will improve progressively as the sequence is finished, as further confirmatory evidence becomes available (particularly from other vertebrate genome sequences, such as those of mouse and *T. nigroviridis*), and as computational methods improve. We intend to create and release updated versions of the IGI and IPI regularly, until they converge to a final accurate list of every human gene. The gene predictions will be linked to RefSeq, HUGO and SWISSPROT identifiers where available, and tracking identifiers between versions will be included, so that individual genes under study can be traced forwards as the human sequence is completed.

Comparative proteome analysis

Knowledge of the human proteome will provide unprecedented opportunities for studies of human gene function. Often clues will be provided by sequence similarity with proteins of known function in model organisms. Such initial observations must then be followed up by detailed studies to establish the actual function of these molecules in humans.

For example, 35 proteins are known to be involved in the vacuolar protein-sorting machinery in yeast. Human genes encoding homologues can be found in the draft human sequence for 34 of these yeast proteins, but precise relationships are not always clear. In nine cases there appears to be a single clear human orthologue (a gene that arose as a consequence of speciation); in 12 cases there are matches to a family of human paralogues (genes that arose owing to intra-genome duplication); and in 13 cases there are matches to specific protein domains^{311–314}. Hundreds of similar stories emerge from the draft sequence, but each merits a detailed interpretation in context. To treat these subjects properly, there will be many following studies, the first of which appear in accompanying papers^{315–323}.

Here, we aim to take a more global perspective on the content of the human proteome by comparing it with the proteomes of yeast, worm, fly and mustard weed. Such comparisons shed useful light on the commonalities and differences among these eukaryotes^{294,324,325}. The analysis is necessarily preliminary, because of the imperfect nature of the human sequence, uncertainties in the gene and protein sets for all of the multicellular organisms considered and our incomplete knowledge of protein structures. Nonetheless, some general patterns emerge. These include insights into fundamental

mechanisms that create functional diversity, including invention of protein domains, expansion of protein and domain families, evolution of new protein architectures and horizontal transfer of genes. Other mechanisms, such as alternative splicing, post-translational modification and complex regulatory networks, are also crucial in generating diversity but are much harder to discern from the primary sequence. We will not attempt to consider the effects of alternative splicing on proteins; we will consider only a single splice form from each gene in the various organisms, even when multiple splice forms are known.

Functional and evolutionary classification. We began by classifying the human proteome on the basis of functional categories and evolutionary conservation. We used the InterPro annotation protocol to identify conserved biochemical and cellular processes. InterPro is a tool for combining sequence-pattern information from four databases. The first two databases (PRINTS³²⁶ and Prosite³²⁷) primarily contain information about motifs corresponding to specific family subtypes, such as type II receptor tyrosine kinases (RTK-II) in particular or tyrosine kinases in general. The second two databases (Pfam³⁰⁷ and Prosite Profile³²⁷) contain information (in the form of profiles or HMMs) about families of structural domains—for example, protein kinase domains. InterPro integrates the motif and domain assignments into a hierarchical classification system; so a protein might be classified at the most detailed level as being an RTK-II, at a more general level as being a kinase specific for tyrosine, and at a still more general level as being a protein kinase. The complete hierarchy of InterPro entries is described at <http://www.ebi.ac.uk/interpro/>. We collapsed the InterPro entries into 12 broad categories, each reflecting a set of cellular functions.

The InterPro families are partly the product of human judgement and reflect the current state of biological and evolutionary knowledge. The system is a valuable way to gain insight into large collections of proteins, but not all proteins can be classified at present. The proportions of the yeast, worm, fly and mustard weed protein sets that are assigned to at least one InterPro family is, for each organism, about 50% (Table 23; refs 307, 326, 327).

About 40% of the predicted human proteins in the IPI could be assigned to InterPro entries and functional categories. On the basis of these assignments, we could compare organisms according to the number of proteins in each category (Fig. 37). Compared with the two invertebrates, humans appear to have many proteins involved in cytoskeleton, defence and immunity, and transcription and translation. These expansions are clearly related to aspects of vertebrate physiology. Humans also have many more proteins that are classified as falling into more than one functional category (426 in human versus 80 in worm and 57 in fly, data not shown). Interestingly, 32% of these are transmembrane receptors.

We obtained further insight into the evolutionary conservation of proteins by comparing each sequence to the complete nonredundant database of protein sequences maintained at NCBI, using the BLASTP computer program³²⁸ and then breaking down the matches according to organismal taxonomy (Fig. 38). Overall, 74% of the proteins had significant matches to known proteins.

Such classifications are based on the presence of clearly detectable homologues in existing databases. Many of these genes have surely evolved from genes that were present in common ancestors but have since diverged substantially. Indeed, one can detect more distant relationships by using sensitive computer programs that can recognize weakly conserved features. Using PSI-BLAST, we can recognize probable nonvertebrate homologues for about 45% of the 'vertebrate-specific' set. Nonetheless, the classification is useful for gaining insights into the commonalities and differences among the proteomes of different organisms.

Probable horizontal transfer. An interesting category is a set of 223 proteins that have significant similarity to proteins from bacteria, but no comparable similarity to proteins from yeast, worm, fly and

Table 23 Properties of genome and proteome in essentially completed eukaryotic proteomes

	Human	Fly	Worm	Yeast	Mustard weed
Number of identified genes	~32,000*	13,338	18,266	6,144	25,706
% with InterPro matches	51	56	50	50	52
Number of annotated domain families	1,262	1,035	1,014	851	1,010
Number of InterPro entries per gene	0.53	0.84	0.63	0.6	0.62
Number of distinct domain architectures	1,695	1,036	1,018	310	—
Percentage of 1-1-1-1	1.40	4.20	3.10	9.20	—
% Signal sequences	20	20	24	11	—
% Transmembrane proteins	20	25	28	15	—
% Repeat-containing	10	11	9	5	—
% Coiled-coil	11	13	10	9	—

The numbers of distinct architectures were calculated using SMART³³⁹ and the percentages of repeat-containing proteins were estimated using Prospero³⁴² and a *P*-value threshold of 10^{-5} . The protein sets used in the analysis were taken from <http://www.ebi.ac.uk/proteome/> for yeast, worm and fly. The proteins from mustard weed were taken from the TAIR website (<http://www.arabidopsis.org/>) on 5 September 2000. The protein set was searched against the InterPro database (<http://www.ebi.ac.uk/interpro/>) using the InterProScan software. Comparison of protein sequences with the InterPro database allows prediction of protein families, domain and repeat families and sequence motifs. The searches used Pfam release 5.2³⁰⁷, Prints release 26.1³²⁶, Prosite release 16³²⁷ and Prosite preliminary profiles. InterPro analysis results are available as Supplementary Information. The fraction of 1-1-1-1 is the percentage of the genome that falls into orthologous groups composed of only one member each in human, fly, worm and yeast.

*The gene number for the human is still uncertain (see text). Table is based on 31,778 known genes and gene predictions.

mustard weed, or indeed from any other (nonvertebrate) eukaryote. These sequences should not represent bacterial contamination in the draft human sequence, because we filtered the sequence to eliminate sequences that were essentially identical to known bacterial plasmid, transposon or chromosomal DNA (such as the host strains for the large-insert clones). To investigate whether these were genuine human sequences, we designed PCR primers for 35 of these genes and confirmed that most could be readily detected directly in human genomic DNA (Table 24). Orthologues of many of these genes have also been detected in other vertebrates (Table 24).

A more detailed computational analysis indicated that at least 113 of these genes are widespread among bacteria, but, among eukaryotes, appear to be present only in vertebrates. It is possible that the genes encoding these proteins were present in both early prokaryotes and eukaryotes, but were lost in each of the lineages of yeast, worm, fly, mustard weed and, possibly, from other nonvertebrate eukaryote lineages. A more parsimonious explanation is that these genes entered the vertebrate (or prevertebrate) lineage by horizontal transfer from bacteria. Many of these genes contain introns, which presumably were acquired after the putative horizontal transfer event. Similar observations indicating probable lineage-specific horizontal gene transfers, as well as intron insertion in the acquired genes, have been made in the worm genome³²⁹.

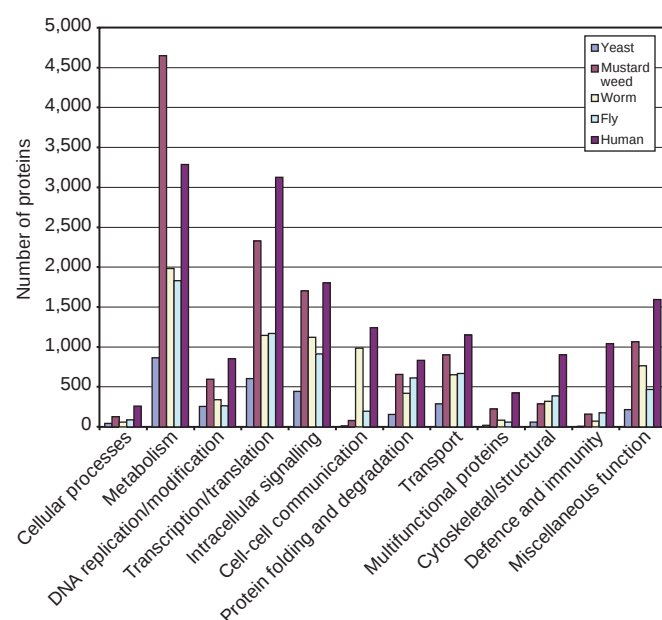


Figure 37 Functional categories in eukaryotic proteomes. The classification categories were derived from functional classification systems, including the top-level biological function category of the Gene Ontology project (GO; see <http://www.geneontology.org>).

We cannot formally exclude the possibility that gene transfer occurred in the opposite direction—that is, that the genes were invented in the vertebrate lineage and then transferred to bacteria. However, we consider this less likely. Under this scenario, the broad distribution of these genes among bacteria would require extensive horizontal dissemination after their initial acquisition. In addition, the functional repertoire of these genes, which largely encode intracellular enzymes (Table 24), is uncharacteristic of vertebrate-specific evolutionary innovations (which appear to be primarily extracellular proteins; see below).

We did not identify a strongly preferred bacterial source for the putative horizontally transferred genes, indicating the likelihood of multiple independent gene transfers from different bacteria (Table 24). Notably, several of the probable recent acquisitions have established (or likely) roles in metabolism of xenobiotics or stress response. These include several hydrolases of different specificities, including epoxide hydrolase, and several dehydrogenases (Table 24). Of particular interest is the presence of two paralogues of monoamine oxidase (MAO), an enzyme of the mitochondrial outer membrane that is central in the metabolism of neuromediators and is a target of important psychiatric drugs^{330–333}. This example shows that at least some of the genes thought to be horizontally transferred into the vertebrate lineage appear to be involved in important physiological functions and so probably have been fixed and maintained during evolution because

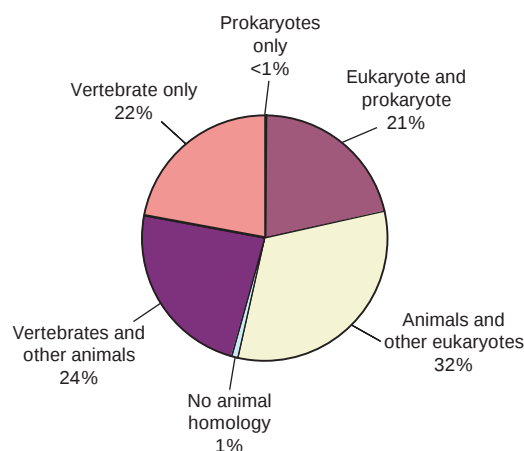


Figure 38 Distribution of the homologues of the predicted human proteins. For each protein, a homologue to a phylogenetic lineage was considered present if a search of the NCBI nonredundant protein sequence database, using the gapped BLASTP program, gave a random expectation (*E*) value of ≤ 0.001 . Additional searches for probable homologues with lower sequence conservation were performed using the PSI-BLAST program, run for three iterations using the same cut-off for inclusion of sequences into the profile³²⁸.

of the increased selective advantage(s) they provide.

Genes shared with fly, worm and yeast. IPI.1 contains apparent homologues of 61% of the fly proteome, 43% of the worm proteome and 46% of the yeast proteome. We next considered the groups of proteins containing likely orthologues and paralogues (genes that arose from intragenome duplication) in human, fly, worm and yeast.

Briefly, we performed all-against-all sequence comparison³³⁴ for the combined protein sets of human, yeast, fly and worm. Pairs of sequences that were one another's best matches in their respective genomes were considered to be potential orthologues. These were then used to identify orthologous groups across three organisms³³⁵. Recent species-specific paralogues were defined by using the all-against-all sequence comparison to cluster the protein set for each organism. For each sequence found in an orthologous group, the recent paralogues were defined to be the largest species-specific cluster including it. The set of paralogues may be inflated by unrecognized splice variants and by fragmentation.

We identified 1,308 groups of proteins, each containing at least one predicted orthologue in each species and many containing additional paralogues. The 1,308 groups contained 3,129 human proteins, 1,445 fly proteins, 1,503 worm proteins and 1,441 yeast proteins. These 1,308 groups represent a conserved core of proteins that are mostly responsible for the basic 'housekeeping' functions of the cell, including metabolism, DNA replication and repair, and translation.

In 564 of the 1,308 groups, one orthologue (and no additional paralogues) could be unambiguously assigned for each of human, fly, worm and yeast. These groups will be referred to as 1-1-1-1 groups. More than half (305) of these groups could be assigned to the functional categories shown in Fig. 37. Within these functional categories, the numbers of groups containing single orthologues in each of the four proteomes was: 19 for cellular processes, 66 for metabolism, 31 for DNA replication and modification, 106 for transcription/translation, 13 for intracellular signalling, 24 for protein folding and degradation, 38 for transport, 5 for

Table 24 Probable vertebrate-specific acquisitions of bacterial genes

Human protein (accession)	Predicted function	Known orthologues in other vertebrates	Bacterial homologues		Human origin confirmed by PCR
			Range	Best hit	
AAG01853.1	Formiminotransferase cyclodeaminase	Pig, rat, chicken	<i>Thermotoga</i> , <i>Thermoplasma</i> , <i>Methylobacter</i>	<i>Thermotoga maritima</i>	Yes
CAB81772.1 AAB59448.1	Na/glucose cotransporter	Rodents, ungulates	Most bacteria	<i>Vibrio parahaemolyticus</i>	Yes (CAB81772, AAC41747.1) NT* (AAB59448.1, AAA36608.1)
AAA36608.1 AAC41747.1 BAA1143.21	Epoxide hydrolase (α/β -hydrolase)	Mouse, <i>Danio</i> , fugu fish	Most bacteria	<i>Pseudomonas aeruginosa</i>	Yes
CAB59628.1 BAA91273.1	Protein-methionine-S-oxide reductase Hypertension-associated protein SA/acetate-CoA ligase	Cow Mouse, rat, cow	Most bacteria Most bacteria	<i>Synechocystis</i> sp. <i>Bacillus halodurans</i>	Yes NT*
CAA75608.1	Glucose-6-phosphate transporter/glycogen storage disease type 1b protein	Mouse, rat	Most bacteria	<i>Chlamydomonas reinhardtii</i>	Yes
AAA59548.1 AAB27229.1 AAF12736.1	Monoamine oxidase Acyl-CoA dehydrogenase, mitochondrial protein	Cow, rat, salmon Mouse, rat, pig	Most bacteria Most bacteria	<i>Mycobacterium tuberculosis</i> <i>P. aeruginosa</i>	Yes Yes
AAA51565.1 IGL_M1_ctg19153_147	Aldose-1-epimerase	Pig (also found in plants)	<i>Streptomyces</i> , <i>Bacillus</i>	<i>Streptomyces coelicolor</i>	Yes
BAA92632.1	Predicted carboxylase (C-terminal domain, N-terminal domain unique)	None	<i>Streptomyces</i> , <i>Rhizobium</i> , <i>Bacillus</i>	<i>S. coelicolor</i>	Yes
BAA34458.1 AAF24044.1 BAA34458.1 BAA91839.1	Uncharacterized protein Uncharacterized protein β -Lactamase superfamily hydrolase Oxidoreductase (Rossmann fold) fused to a six-transmembrane protein	None None None None (several human paralogues of both parts)	Gamma-proteobacteria Most bacteria Most bacteria <i>Actinomyces</i> , <i>Leptospira</i> ; more distant homologues in other bacteria	<i>Escherichia coli</i> <i>T. maritima</i> <i>Synechocystis</i> sp. <i>S. coelicolor</i>	Yes Yes Yes Yes
BAA92073.1 BAA92133.1	Oxidoreductase (Rossmann fold) α/β -hydrolase	None None	<i>Synechocystis</i> , <i>Pseudomonas</i> <i>Rickettsia</i> ; more distant homologues in other bacteria	<i>Synechocystis</i> sp. <i>Rickettsia prowazekii</i>	Yes Yes
BAA91174.1	ADP-ribosylglycohydrolase	None	<i>Streptomyces</i> , <i>Aquifex</i> , <i>Archaeoglobus</i> (archaeon), <i>E. coli</i>	<i>S. coelicolor</i>	Yes
AAA60043.1	Thymidine phosphorylase/endothelial cell growth factor	None	Most bacteria	<i>Bacillus stearothermophilus</i>	Yes
BAA86552.1	Ribosomal protein S6-glutamic acid ligase	None	Most bacteria and archaea	<i>Haemophilus influenzae</i>	Yes
IGL_M1_ctg12741_7 IGL_M1_ctg13238_61	Ribosomal protein S6-glutamic acid ligase (paralogue of the above) Hydratase	None None	Most bacteria and archaea <i>Synechocystis</i> , <i>Sphingomonas</i>	<i>H. influenzae</i> <i>Synechocystis</i> sp.	Yes Yes
IGL_M1_ctg13305_116	Homologue of histone macro-2A C-terminal domain, predicted phosphatase	None (several human paralogues, RNA viruses)	<i>Thermotoga</i> , <i>Alcaligenes</i> , <i>E. coli</i> , more distant homologues in other bacteria	<i>T. maritima</i>	Yes
IGL_M1_ctg14420_10 IGL_M1_ctg16010_18 IGL_M1_ctg16227_58 IGL_M1_ctg25107_24	Sugar transporter Predicted metal-binding protein Pseudouridine synthase Surfactin synthetase domain	None None None None	Most bacteria Most bacteria Most bacteria Gram-positive bacteria, <i>Actinomyces</i> , <i>Cyanobacteria</i>	<i>Synechocystis</i> sp. <i>Borrelia burgdorferi</i> <i>Zymomonas mobilis</i> <i>Bacillus subtilis</i>	Yes Yes Yes Yes

* NT, not tested.

Representative genes confirmed by PCR to be present in the human genome. The similarity to a bacterial homologue was considered to be 'significantly' greater than that to eukaryotic homologues if the difference in alignment scores returned by BLASTP was greater than 30 bits (~9 orders of magnitude in terms of E-value). A complete, classified and annotated list of probable vertebrate-specific horizontal gene transfers detected in this analysis is available as Supplementary Information. cDNA sequences for each protein were searched, using the SSAHA algorithm, against the draft genome sequence. Primers were designed and PCR was performed using three human genomic samples and a random BAC clone. The predicted genes were considered to be present in the human genome if a band of the expected size was found in all three human samples but not in the control clone.

multifunctional proteins and 3 for cytoskeletal/structural. No such groups were found for defence and immunity or cell–cell communication.

The 1-1-1 groups probably represent key functions that have not undergone duplication and elaboration in the various lineages. They include many anabolic enzymes responsible for such functions as respiratory chain and nucleotide biosynthesis. In contrast, there are few catabolic enzymes. As anabolic pathways branch less frequently than catabolic pathways, this indicates that alternative routes and displacements are more frequent in catabolic reactions. If proteins from the single-celled yeast are excluded from the analysis, there are 1,195 1-1-1 groups. The additional groups include many examples of more complex signalling proteins, such as receptor-type and src-like tyrosine kinases, likely to have arisen early in the metazoan lineage. The fact that this set comprises only a small proportion of the proteome of each of the animals indicates that, apart from a modest conserved core, there has been extensive elaboration and innovation within the protein complement.

Most proteins do not show simple 1-1-1 orthologous relationships across the three animals. To illustrate this, we investigated the nuclear hormone receptor family. In the human proteome, this family consists of 60 different ‘classical’ members, each with a zinc finger and a ligand-binding domain. In comparison, the fly proteome has 19 and the worm proteome has 220. As shown in Fig. 39, few simple orthologous relationships can be derived among these homologues. And, where potential subgroups of orthologues and

paralogues could be identified, it was apparent that the functions of the subgroup members could differ significantly. For example, the fly receptor for the fly-specific hormone ecdysone and the human retinoic acid receptors cluster together on the basis of sequence similarity. Such examples underscore that the assignment of functional similarity on the basis of sequence similarities among these three organisms is not trivial in most cases.

New vertebrate domains and proteins. We then explored how the proteome of vertebrates (as represented by the human) differs from those of the other species considered. The 1,262 InterPro families were scanned to identify those that contain only vertebrate proteins. Only 94 (7%) of the families were ‘vertebrate-specific’. These represent 70 protein families and 24 domain families. Only one of the 94 families represents enzymes, which is consistent with the ancient origins of most enzymes³³⁶. The single vertebrate-specific enzyme family identified was the pancreatic or eosinophil-associated ribonucleases. These enzymes evolved rapidly, possibly to combat vertebrate pathogens³³⁷.

The relatively small proportion of vertebrate-specific multicopy families suggests that few new protein domains have been invented in the vertebrate lineage, and that most protein domains trace at least as far back as a common animal ancestor. This conclusion must be tempered by the fact that the InterPro classification system is incomplete; additional vertebrate-specific families undoubtedly exist that have not yet been recognized in the InterPro system.

The 94 vertebrate-specific families appear to reflect important physiological differences between vertebrates and other eukaryotes. Defence and immunity proteins (23 families) and proteins that function in the nervous system (17 families) are particularly enriched in this set. These data indicate the recent emergence or rapid divergence of these proteins.

Representative human proteins were previously known for nearly all of the vertebrate-specific families. This was not surprising, given the anthropocentrism of biological research. However, the analysis did identify the first mammalian proteins belonging to two of these families. Both of these families were originally defined in fish. The first is the family of polar fish antifreeze III proteins. We found a human sialic acid synthase containing a domain homologous to polar fish antifreeze III protein (BAA91818.1). This finding suggests that fish created the antifreeze function by adaptation of this domain. We also found a human protein (CAB60269.1) homologous to the ependymin found in teleost fish. Ependymins are major glycoproteins of fish brains that have been claimed to be involved in long-term memory formation³³⁸. The function of the mammalian ependymin homologue will need to be elucidated.

New architectures from old domains. Whereas there appears to be only modest invention at the level of new vertebrate protein domains, there appears to be substantial innovation in the creation of new vertebrate proteins. This innovation is evident at the level of domain architecture, defined as the linear arrangement of domains within a polypeptide. New architectures can be created by shuffling, adding or deleting domains, resulting in new proteins from old parts.

We quantified the number of distinct protein architectures found in yeast, worm, fly and human by using the SMART annotation resource³³⁹ (Fig. 40). The human proteome set contained 1.8 times as many protein architectures as worm or fly and 5.8 times as many as yeast. This difference is most prominent in the recent evolution of novel extracellular and transmembrane architectures in the human lineage. Human extracellular proteins show the greatest innovation: the human has 2.3 times as many extracellular architectures as fly and 2.0 times as many as worm. The larger number of human architectures does not simply reflect differences in the number of domains known in these organisms; the result remains qualitatively the same even if the number of architectures in each organism is normalized by dividing by the total number of domains (not shown). (We also checked that the larger number of human

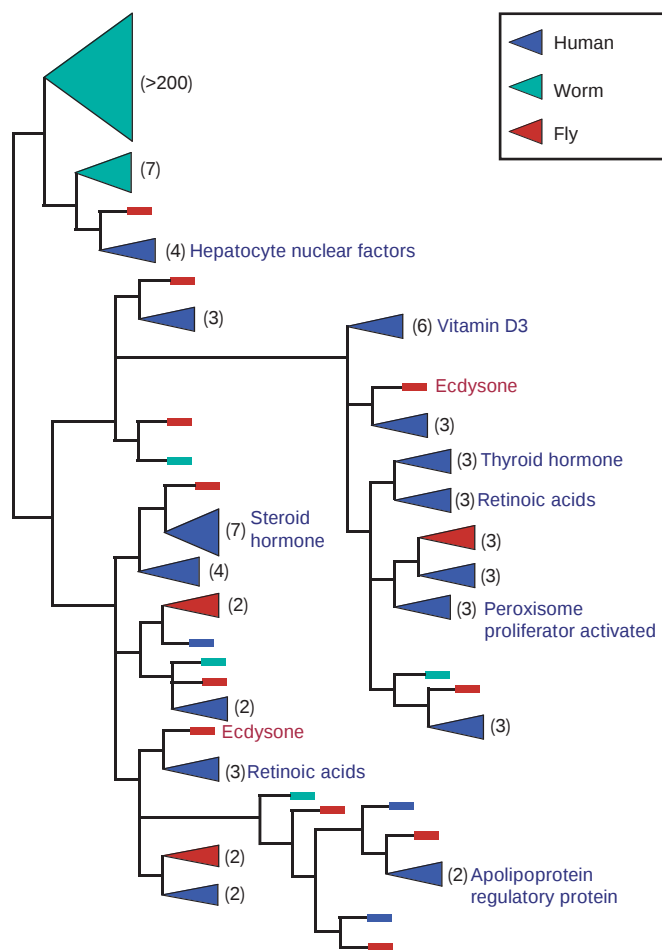


Figure 39 Simplified cladogram (relationship tree) of the ‘many-to-many’ relationships of classical nuclear receptors. Triangles indicate expansion within one lineage; bars represent single members. Numbers in parentheses indicate the number of paralogues in each group.

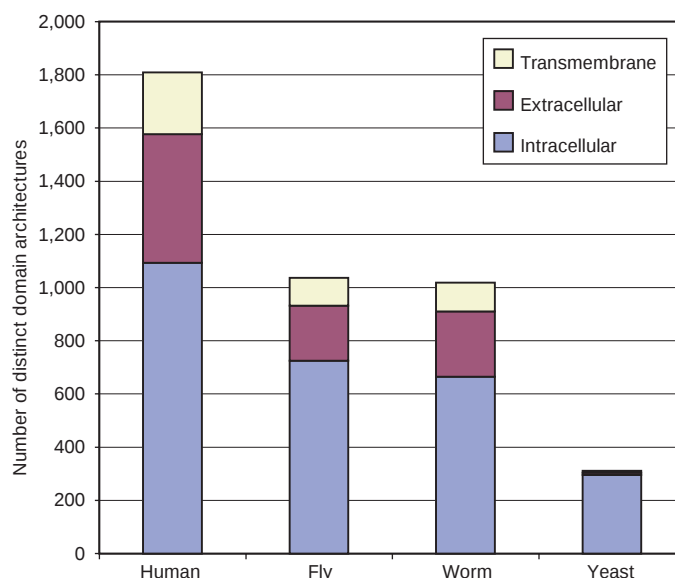


Figure 40 Number of distinct domain architectures in the four eukaryotic genomes, predicted using SMART³³⁹. The number of architectures is split into three cellular environments: intracellular, extracellular and membrane-associated. The increase in architectures for the human, relative to the other lineages, is seen when these numbers

are normalized with respect to the numbers of domains predicted in each phylum. To avoid artefactual results from the relatively low detection rate for some repeat types, tandem occurrences of tetratricopeptide, armadillo, EF-hand, leucine-rich, WD40 or ankyrin repeats or C2H2-type zinc fingers were treated as single occurrences.

architectures could not be an artefact resulting from erroneous gene predictions. Three-quarters of the architectures can be found in known genes, which already yields an increase of about 50% over worm and fly. We expect the final number of human architectures to grow as the complete gene set is identified.)

A related measure of proteome complexity can be obtained by considering an individual domain and counting the number of different domain types with which it co-occurs. For example, the trypsin-like serine protease domain (number 12 in Fig. 41) co-occurs with 18 domain types in human (including proteins involved in the mammalian complement system, blood coagulation, and fibrinolytic and related systems). By contrast, the trypsin-like serine protease domain occurs with only eight other domains in fly, five in worm and one in yeast. Similar results for 27 common domains are shown in Fig. 41. In general, there are more different co-occurring domains in the human proteome than in the other proteomes.

One mechanism by which architectures evolve is through the fusion of additional domains, often at one or both ends of the proteins. Such 'domain accretion'³⁴⁰ is seen in many human proteins when compared with proteins from other eukaryotes. The effect is illustrated by several chromatin-associated proteins (Fig. 42). In these examples, the domain architectures of human proteins differ from those found in yeast, worm and fly proteins only by the addition of domains at their termini.

Among chromatin-associated proteins and transcription factors, a significant proportion of domain architectures is shared between the vertebrate and fly, but not with worm (Fig. 43a). The trend was even more prominent in architectures of proteins involved in another key cellular process, programmed cell death (Fig. 43b). These examples might seem to bear upon the unresolved issue of the evolutionary branching order of worms, flies and humans, suggesting that worms branched off first. However, there were other cases in which worms and humans shared architectures not present in fly. A global analysis of shared architectures could not conclusively distinguish between the two models, given the possibility of lineage-specific loss of architectures. Comparison of protein architectures may help to resolve the evolutionary issue, but it will require more detailed analyses of many protein families.

New physiology from old proteins. An important aspect of

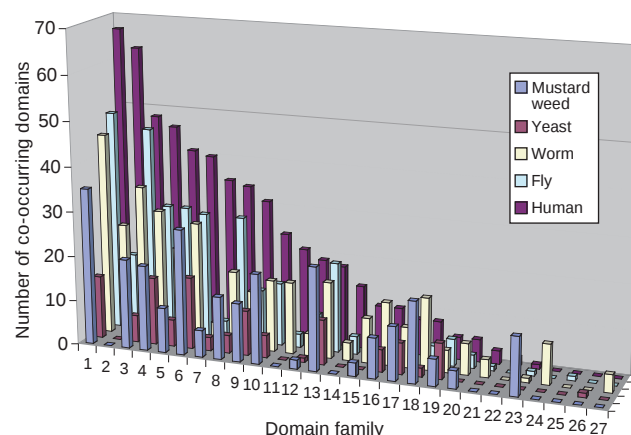


Figure 41 Number of different Pfam domain types that co-occur in the same protein, for each of the 10 most common domain families in each of the five eukaryotic proteomes. Because some common domain families are shared, there are 27 families rather than 50. The data are ranked according to decreasing numbers of human co-occurring Pfam domains. The domain families are: (1) eukaryotic protein kinase [IPR000719]; (2) immunoglobulin domain [IPR003006]; (3) ankyrin repeat [IPR002110]; (4) RING finger [IPR001841]; (5) C2H2-type zinc finger [IPR000822]; (6) ATP/GTP-binding P-loop [IPR001687]; (7) reverse transcriptase (RNA-dependent DNA polymerase) [IPR000477]; (8) leucine-rich repeat [IPR001611]; (9) G-protein β WD-40 repeats [IPR001680]; (10) RNA-binding region RNP-1 (RNA recognition motif) [IPR000504]; (11) C-type lectin domain [IPR001304]; (12) serine proteases, trypsin family [IPR001254]; (13) helicase C-terminal domain [IPR001650]; (14) collagen triple helix repeat [IPR000087]; (15) rhodopsin-like GPCR superfamily [IPR000276]; (16) esterase/lipase/thioesterase [IPR000379]; (17) Myb DNA-binding domain [IPR001005]; (18) F-box domain [IPR001810]; (19) ATP-binding transport protein, 2nd P-loop motif [IPR001051]; (20) homeobox domain [IPR001356]; (21) C4-type steroid receptor zinc finger [IPR001628]; (22) sugar transporter [IPR001066]; (23) PPR repeats [IPR002885]; (24) seven-helix G-protein-coupled receptor, worm (probably olfactory) family [IPR000168]; (25) cytochrome P450 enzyme [IPR001128]; (26) fungal transcriptional regulatory protein, N terminus [IPR001138]; (27) domain of unknown function DUF38 [IPR002900].

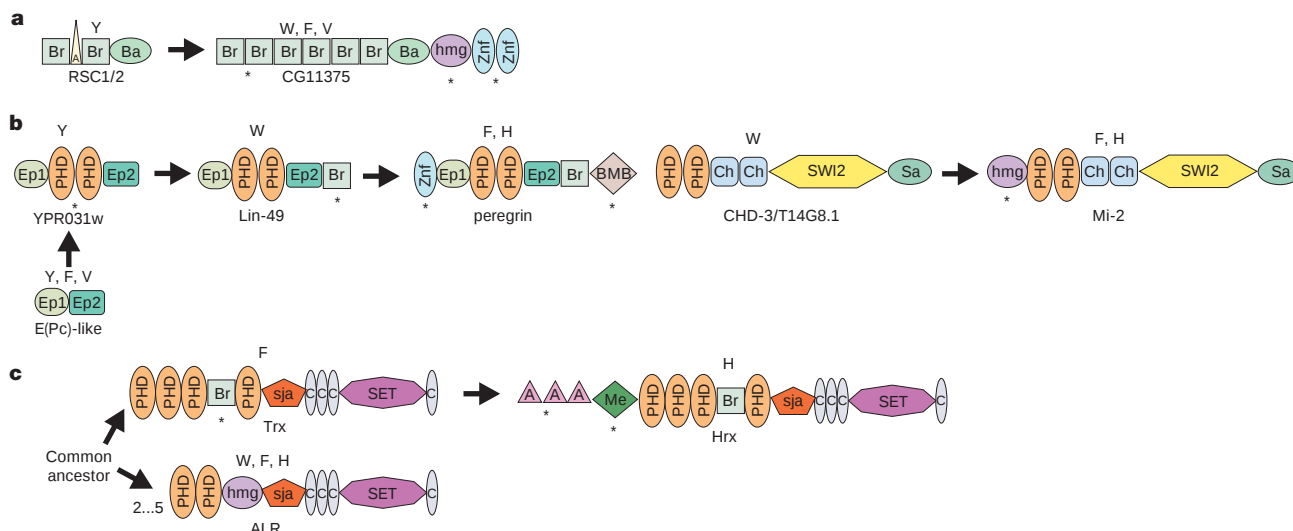


Figure 42 Examples of domain accretion in chromatin proteins. Domain accretion in various lineages before the animal divergence, in the apparent coelomate lineage and the vertebrate lineage are shown using schematic representations of domain architectures (not to scale). Asterisks, mobile domains that have participated in the accretion. Species in which a domain architecture has been identified are indicated above the diagram (Y, yeast; W, worm; F, fly; V, vertebrate). Protein names are below the diagrams. The

domains are SET, a chromatin protein methyltransferase domain; SWI2, a superfamily II helicase/ATPase domain; Sa, sant domain; Br, bromo domain; Ch, chromodomain; C, a cysteine triad motif associated with the Msl-2 and SET domains; A, AT hook motif; EP1/EP2, enhancer of polycomb domains 1 and 2; Znf, zinc finger; sja, SET-JOR-associated domain (L. Aravind, unpublished); Me, DNA methylase/Hrx-associated DNA binding zinc finger; Ba, bromo-associated homology motif. **a–c**, Different examples of accretion.

vertebrate innovation lies in the expansion of protein families. Table 25 shows the most prevalent protein domains and protein families in humans, together with their relative ranks in the other species. About 60% of families are more numerous in the human than in any of the other four organisms. This shows that gene duplication has been a major evolutionary force during vertebrate evolution. A

comparison of relative expansions in human versus fly is shown in Fig. 44.

Many of the families that are expanded in human relative to fly and worm are involved in distinctive aspects of vertebrate physiology. An example is the family of immunoglobulin (IG) domains, first identified in antibodies thirty years ago. Classic (as opposed to divergent) IG domains are completely absent from the yeast and mustard weed proteomes and, although prokaryotic homologues exist, they have probably been transferred horizontally from metazoans³⁴¹. Most IG superfamily proteins in invertebrates are cell-surface proteins. In vertebrates, the IG repertoire includes immune functions such as those of antibodies, MHC proteins, antibody receptors and many lymphocyte cell-surface proteins. The large expansion of IG domains in vertebrates shows the versatility of a single family in evoking rapid and effective response to infection.

Two prominent families are involved in the control of development. The human genome contains 30 fibroblast growth factors (FGFs), as opposed to two FGFs each in the fly and worm. It contains 42 transforming growth factor- β s (TGF β s) compared with nine and six in the fly and worm, respectively. These growth factors are involved in organogenesis, such as that of the liver and the lung. A fly FGF protein, branchless, is involved in developing respiratory organs (tracheae) in embryos³⁴². Thus, developmental triggers of morphogenesis in vertebrates have evolved from related but simpler systems in invertebrates³⁴³.

Another example is the family of intermediate filament proteins, with 127 family members. This expansion is almost entirely due to 111 keratins, which are chordate-specific intermediate filament proteins that form filaments in epithelia. The large number of human keratins suggests multiple cellular structural support roles for the many specialized epithelia of vertebrates.

Finally, the olfactory receptor genes comprise a huge gene family of about 1,000 genes and pseudogenes^{344,345}. The number of olfactory receptors testifies to the importance of the sense of smell in vertebrates. A total of 906 olfactory receptor genes and pseudogenes could be identified in the draft genome sequence, two-thirds of which were not previously annotated. About 80% are found in about two dozen clusters ranging from 6 to 138 genes and encompassing about 30 Mb (~1%) of the human genome. Despite the importance of smell among our vertebrate ancestors, hominids

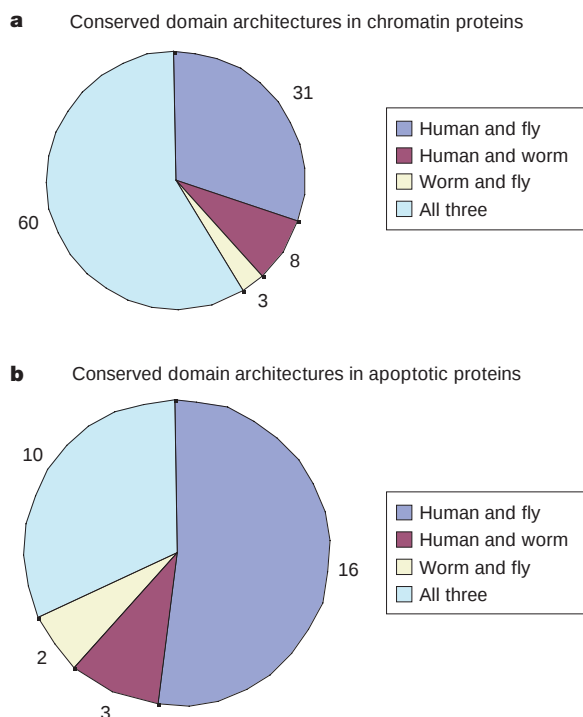


Figure 43 Conservation of architectures between animal species. The pie charts illustrate the shared domain architectures of apparent orthologues that are conserved in at least two of the three sequenced animal genomes. If an architecture was detected in fungi or plants, as well as two of the animal lineages, it was omitted as ancient and its absence in the third animal lineage attributed to gene loss. **a**, Chromatin-associated proteins. **b**, Components of the programmed cell death system.

appear to have considerably less interest in this sense. About 60% of the olfactory receptors in the draft genome sequence have disrupted ORFs and appear to be pseudogenes, consistent with recent reports^{344,346} suggesting massive functional gene loss in the last 10 Myr^{347,348}. Interestingly, there appears to be a much higher proportion of intact genes among class I than class II olfactory receptors, suggesting functional importance.

Vertebrates are not unique in employing gene family expansion. For many domain types, expansions appear to have occurred independently in each of the major eukaryotic lineages. A good example is the classical C2H2 family of zinc finger domains, which have expanded independently in the yeast, worm, fly and human lineages (Fig. 45). These independent expansions have resulted in numerous C2H2 zinc finger domain-containing proteins that are specific to each lineage. In flies, the important components of the C2H2 zinc finger expansion are architectures in which it is combined with the POZ domain and the C4DM domain (a metal-binding domain found only in fly). In humans, the most prevalent expansions are combinations of the C2H2 zinc finger with POZ (independent of the one in insects) and the vertebrate-specific KRAB and SCAN domains.

The homeodomain is similarly expanded in all animals and is present in both architectures that are conserved and lineage-specific architectures (Fig. 45). This indicates that the ancestral animal probably encoded a significant number of homeodomain proteins, but subsequent evolution involved multiple, independent expansions and domain shuffling after lineages diverged. Thus, the most prevalent transcription factor families are different in worm, fly and human (Fig. 45). This has major biological implications because transcription factors are critical in animal development and differ-

entiation. The emergence of major variations in the developmental body plans that accompanied the early radiation of the animals³⁴⁹ could have been driven by lineage-specific proliferation of such transcription factors. Beyond these large expansions of protein families, protein components of particular functional systems such as the cell death signalling system show a general increase in diversity and numbers in the vertebrates relative to other animals. For example, there are greater numbers of and more novel architectures in cell death regulatory proteins such as BCL-2, TNFR and NFκB from vertebrates.

Conclusion. Five lines of evidence point to an increase in the complexity of the proteome from the single-celled yeast to the multicellular invertebrates and to vertebrates such as the human. Specifically, the human contains greater numbers of genes, domain and protein families, paralogues, multidomain proteins with multiple functions, and domain architectures. According to these measures, the relatively greater complexity of the human proteome is a consequence not simply of its larger size, but also of large-scale protein innovation.

An important question is the extent to which the greater phenotypic complexity of vertebrates can be explained simply by two- or threefold increases in proteome complexity. The real explanation may lie in combinatorial amplification of these modest differences, by mechanisms that include alternative splicing, post-translational modification and cellular regulatory networks. The potential numbers of different proteins and protein-protein interactions are vast, and their actual numbers cannot readily be discerned from the genome sequence. Elucidating such system-level properties presents one of the great challenges for modern biology.

Table 25 The most populous InterPro families in the human proteome and other species

InterPro ID	Human		Fly		Worm		Yeast		Mustard weed		
	No. of genes	Rank	No. of genes	Rank	No. of genes	Rank	No. of genes	Rank	No. of genes	Rank	
IPR003006	765	(1)	140	(9)	64	(34)	0	(na)	0	(na)	Immunoglobulin domain
PR000822	706	(2)	357	(1)	151	(10)	48	(7)	115	(20)	C2H2 zinc finger
IPR000719	575	(3)	319	(2)	437	(2)	121	(1)	1049	(1)	Eukaryotic protein kinase
IPR000276	569	(4)	97	(14)	358	(3)	0	(na)	16	(84)	Rhodopsin-like GPCR superfamily
IPR001687	433	(5)	198	(4)	183	(7)	97	(2)	331	(5)	P-loop motif
IPR000477	350	(6)	10	(65)	50	(41)	6	(36)	80	(35)	Reverse transcriptase (RNA-dependent DNA polymerase)
IPR000504	300	(7)	157	(6)	96	(21)	54	(6)	255	(8)	rm domain
IPR001680	277	(8)	162	(5)	102	(19)	91	(3)	210	(10)	G-protein β WD-40 repeats
IPR002110	276	(9)	105	(13)	107	(17)	19	(23)	120	(18)	Ankyrin repeat
IPR001356	267	(10)	148	(7)	109	(15)	9	(33)	118	(19)	Homeobox domain
IPR001849	252	(11)	77	(22)	71	(31)	27	(17)	27	(73)	PH domain
IPR002048	242	(12)	111	(12)	81	(25)	15	(27)	167	(12)	EF-hand family
IPR000561	222	(13)	81	(20)	113	(14)	0	(na)	17	(83)	EGF-like domain
IPR001452	215	(14)	72	(23)	62	(35)	25	(18)	3	(97)	SH3 domain
IPR001841	210	(15)	114	(11)	126	(12)	35	(12)	379	(4)	RING finger
IPR001611	188	(16)	115	(10)	54	(38)	7	(35)	392	(2)	Leucine-rich repeat
IPR001909	171	(17)	0	(na)	0	(na)	0	(na)	0	(na)	KRAB box
IPR001777	165	(18)	63	(27)	51	(40)	2	(40)	4	(96)	Fibronectin type III domain
IPR001478	162	(19)	70	(24)	66	(33)	2	(40)	15	(85)	PDZ domain
IPR001650	155	(20)	87	(17)	78	(27)	79	(4)	148	(13)	Helicase C-terminal domain
IPR001440	150	(21)	86	(18)	46	(43)	36	(11)	125	(17)	TPR repeat
IPR002216	133	(22)	65	(26)	99	(20)	2	(40)	31	(69)	Ion transport protein
IPR001092	131	(23)	84	(19)	41	(46)	7	(35)	106	(24)	Helix-loop-helix DNA-binding domain
IPR000008	123	(24)	43	(34)	36	(49)	9	(33)	82	(34)	C2 domain
IPR001664	119	(25)	4	(71)	22	(63)	1	(41)	2	(98)	SH2 domain
IPR001254	118	(26)	210	(3)	12	(73)	1	(41)	15	(85)	Serine protease, trypsin family
IPR002126	114	(27)	19	(56)	16	(69)	0	(na)	0	(na)	Cadherin domain
IPR000210	113	(28)	78	(21)	117	(13)	1	(41)	54	(50)	BTB/POZ domain
IPR000387	112	(29)	35	(40)	108	(16)	12	(30)	21	(79)	Tyrosine-specific protein phosphatase and dual specificity protein phosphatase family
IPR000087	106	(30)	18	(57)	169	(9)	0	(na)	5	(95)	Collagen triple helix repeat
IPR000379	94	(31)	141	(8)	134	(11)	40	(10)	194	(11)	Esterase/lipase/thioesterase
IPR000910	89	(32)	38	(38)	18	(67)	8	(34)	18	(82)	HMG1/2 (high mobility group) box
IPR000130	87	(33)	56	(29)	92	(22)	8	(34)	12	(88)	Neutral zinc metalloprotease
IPR001965	84	(34)	37	(39)	24	(61)	16	(26)	71	(39)	PHD-finger
IPR000636	83	(35)	32	(43)	24	(61)	1	(41)	14	(86)	Cation channels (non-ligand gated)
IPR001781	81	(36)	38	(38)	36	(49)	4	(38)	8	(92)	LIM domain
IPR002035	81	(36)	8	(67)	45	(44)	3	(39)	17	(83)	VWA domain
IPR001715	80	(37)	33	(42)	30	(55)	3	(39)	18	(82)	Calponin homology domain
IPR000198	77	(38)	20	(55)	20	(65)	10	(32)	9	(91)	RhoGAP domain

Forty most populous Interpro families found in the human proteome compared with equivalent numbers from other species. na, not applicable (used when there are no proteins in an organism in that family).

Segmental history of the human genome

In bacteria, genomic segments often convey important information about function: genes located close to one another often encode proteins in a common pathway and are regulated in a common operon. In mammals, genes found close to each other only rarely have common functions, but they are still interesting because they have a common history. In fact, the study of genomic segments can shed light on biological events as long as 500 Myr ago and as recently as 20,000 years ago.

Conserved segments between human and mouse

Humans and mice shared a common ancestor about 100 Myr ago. Despite the 200 Myr of evolutionary distance between the species, a significant fraction of genes show synteny between the two, being preserved within conserved segments. Genes tightly linked in one mammalian species tend to be linked in others. In fact, conserved segments have been observed in even more distant species: humans show conserved segments with fish^{350,351} and even with invertebrates such as fly and worm³⁵². In general, the likelihood that a syntenic relationship will be disrupted correlates with the physical distance between the loci and the evolutionary distance between the species.

Studying conserved segments between human and mouse has several uses. First, conservation of gene order has been used to identify likely orthologues between the species, particularly when investigating disease phenotypes. Second, the study of conserved segments among genomes helps us to deduce evolutionary ancestry.

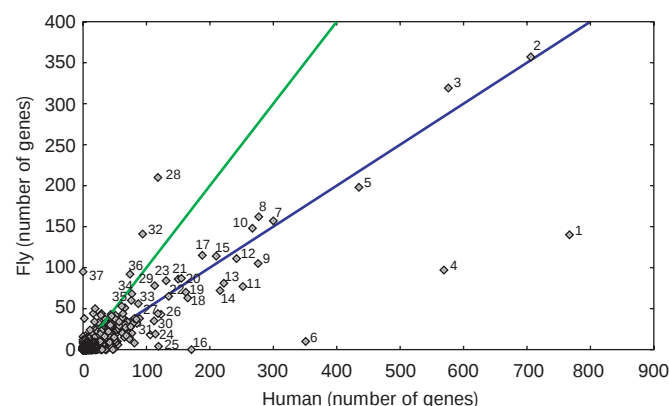


Figure 44 Relative expansions of protein families between human and fly. These data have not been normalized for proteomic size differences. Blue line, equality between normalized family sizes in the two organisms. Green line, equality between unnormalized family sizes. Numbered InterPro entries: (1) immunoglobulin domain [IPR003006]; (2) zinc finger, C2H2 type [IPR00822]; (3) eukaryotic protein kinase [IPR00719]; (4) rhodopsin-like GPCR superfamily [IPR00276]; (5) ATP/GTP-binding site motif A (P-loop) [IPR01687]; (6) reverse transcriptase (RNA-dependent DNA polymerase) [IPR00477]; (7) RNA-binding region RNP-1 (RNA recognition motif) [IPR00504]; (8) G-protein β WD-40 repeats [IPR01680]; (9) ankyrin repeat [IPR02110]; (10) homeobox domain [IPR01356]; (11) PH domain [IPR01849]; (12) EF-hand family [IPR02048]; (13) EGF-like domain [IPR00561]; (14) Src homology 3 (SH3) domain [IPR01452]; (15) RING finger [IPR01841]; (16) KRAB box [IPR01909]; (17) leucine-rich repeat [IPR01611]; (18) fibronectin type III domain [IPR01777]; (19) PDZ domain (also known as DHR or GLGF) [IPR01478]; (20) TPR repeat [IPR01440]; (21) helicase C-terminal domain [IPR01650]; (22) ion transport protein [IPR02216]; (23) helix-loop-helix DNA-binding domain [IPR01092]; (24) cadherin domain [IPR02126]; (25) intermediate filament proteins [IPR01664]; (26) C2 domain [IPR00008]; (27) Src homology 2 (SH2) domain [IPR00980]; (28) serine proteases, trypsin family [IPR01254]; (29) BTB/POZ domain [IPR00210]; (30) tyrosine-specific protein phosphatase and dual specificity protein phosphatase family [IPR00387]; (31) collagen triple helix repeat [IPR00087]; (32) esterase/lipase/thioesterase [IPR00379]; (33) neutral zinc metalloproteases, zinc-binding region [IPR00130]; (34) ATP-binding transport protein, 2nd P-loop motif [IPR01051]; (35) ABC transporters family [IPR01617]; (36) cytochrome P450 enzyme [IPR01128]; (37) insect cuticle protein [IPR00618].

And third, detailed comparative maps may assist in the assembly of the mouse sequence, using the human sequence as a scaffold.

Two types of linkage conservation are commonly described³⁵³. ‘Conserved synteny’ indicates that at least two genes that reside on a common chromosome in one species are also located on a common chromosome in the other species. Syntenic loci are said to lie in a ‘conserved segment’ when not only the chromosomal position but the linear order of the loci has been preserved, without interruption by other chromosomal rearrangements.

An initial survey of homologous loci in human and mouse³⁵⁴ suggested that the total number of conserved segments would be about 180. Subsequent estimates based on increasingly detailed comparative maps have remained close to this projection^{353,355,356} (<http://www.informatics.jax.org>). The distribution of segment lengths has corresponded reasonably well to the truncated negative exponential curve predicted by the random breakage model³⁵⁷.

The availability of a draft human genome sequence allows the first global human–mouse comparison in which human physical distances can be measured in Mb, rather than cM or orthologous gene counts. We identified likely orthologues by reciprocal comparison of the human and mouse mRNAs in the LocusLink database, using megaBLAST. For each orthologous pair, we mapped the location of the human gene in the draft genome sequence and then checked the location of the mouse gene in the Mouse Genome Informatics database (<http://www.informatics.jax.org>). Using a conservative threshold, we identified 3,920 orthologous pairs in which the human gene could be mapped on the draft genome sequence with high confidence. Of these, 2,998 corresponding mouse genes had a known position in the mouse genome. We then searched for definitive conserved segments, defined as human regions containing orthologues of at least two genes from the same mouse chromosome region (< 15 cM) without interruption by segments from other chromosomes.

We identified 183 definitive conserved segments (Fig. 46). The average segment length was 15.4 Mb, with the largest segment being 90.5 Mb and the smallest 24 kb. There were also 141 ‘singletons’, segments that contained only a single locus; these are not counted in the statistics. Although some of these could be short conserved segments, they could also reflect incorrect choices of orthologues or problems with the human or mouse maps. Because of this conservative approach, the observed number of definitive segments is likely to be lower than the correct total. One piece of evidence for this conclusion comes from a more detailed analysis on human chromosome 7 (ref. 358), which identified 20 conserved segments, of which three were singletons. Our analysis revealed only 13 definitive segments on this chromosome, with nine singletons.

The frequency of observing a particular gene count in a conserved segment is plotted on a logarithmic scale in Fig. 47. If chromosomal breaks occur in a random fashion (as has been proposed) and differences in gene density are ignored, a roughly straight line should result. There is a clear excess for $n = 1$, suggesting that 50% or more of the singletons are indeed artefactual. Thus, we estimate that true number of conserved segments is around 190–230, in good agreement with the original Nadeau–Taylor prediction³⁵⁴.

Figure 48 shows a plot of the frequency of lengths of conserved segments, where the x-axis scale is shown in Mb. As before, there is a fair amount of scatter in the data for the larger segments (where the numbers are small), but the trend appears to be consistent with a random breakage model.

We attempted to ascertain whether the breakpoint regions have any special characteristics. This analysis was complicated by imprecision in the positioning of these breaks, which will tend to blur any relationships. With 2,998 orthologues, the average interval within which a break is known to have occurred is about 1.1 Mb. We compared the aggregate features of these breakpoint intervals with the genome as a whole. The mean gene density was lower in breakpoint regions than in the conserved segments (13.8 versus

18.6 per Mb). This suggests that breakpoints may be more likely to occur or to undergo fixation in gene-poor intervals than in gene-rich intervals. The occurrence of breakpoints may be promoted by homologous recombination among repeated sequences³⁵⁹. When the sequence of the mouse genome is finished, this analysis can be revisited more precisely.

A number of examples of extended conserved segments and synteny are apparent in Fig. 46. As has been noted, almost all human genes on chromosome 17 are found on mouse chromosome 11, with two members of the placental lactogen family from mouse 13 inserted. Apart from two singleton loci, human chromosome 20 appears to be entirely orthologous to mouse chromosome 2, apparently in a single segment. The largest apparently contiguous conserved segment in the human genome is on chromosome 4, including roughly 90.5 Mb of human DNA that is orthologous to mouse chromosome 5. This analysis also allows us to infer the likely location of thousands of mouse genes for which the human orthologue has been located in the draft genome sequence but the mouse locus has not yet been mapped.

With about 200 conserved segments between mouse and human and about 100 Myr of evolution from their common ancestor³⁶⁰, we obtain an estimated rate of about 1.0 chromosomal rearrangement being fixed per Myr. However, there is good evidence that the rate of chromosomal rearrangement (like the rate of nucleotide substitutions; see above) differs between the two species. Among mammals, rodents may show unusually rapid chromosome alteration. By comparison, very few rearrangements have been observed among primates, and studies of a broader array of mammalian orders, including cats, cows, sheep and pigs, suggest an average rate of chromosome alteration of only about 0.2 rearrangements per Myr

in these lineages³⁶¹. Additional evidence that rodents are outliers comes from a recent analysis of synteny between the human and zebrafish genomes. From a study of 523 orthologues, it was possible to project 418 conserved segments³⁵⁰. Assuming 400 Myr since a common vertebrate ancestor of zebrafish and humans³⁶², we obtain an estimate of 0.52 rearrangements per Myr. Recent estimates of rearrangement rates in plants have suggested bimodality, with some pairs showing rates of 0.15–0.41 rearrangements per Myr, and others showing higher rates of 1.1–1.3 rearrangements per Myr³⁶³. With additional detailed genome maps of multiple species, it should be possible to determine whether this particular molecular clock is truly operating at a different rate in various branches of the evolutionary tree, and whether variations in that rate are bimodal or continuous. It should also be possible to reconstruct the karyotypes of common ancestors.

Ancient duplicated segments in the human genome

Another approach to genomic history is to study segmental duplications within the human genome. Earlier, we discussed examples of recent duplications of genomic segments to pericentromeric and subtelomeric regions. Most of these events appear to be evolutionary dead-ends resulting in nonfunctional pseudogenes; however, segmental duplication is also an important mode of evolutionary innovation: a duplication permits one copy of each gene to drift and potentially to acquire a new function.

Segmental duplications can occur through unequal crossing over to create gene families in specific chromosomal regions. This mechanism can create both small families, such as the five related genes of the β -globin cluster on chromosome 11, and large ones, such as the olfactory receptor gene clusters, which together contain nearly 1,000 genes and pseudogenes.

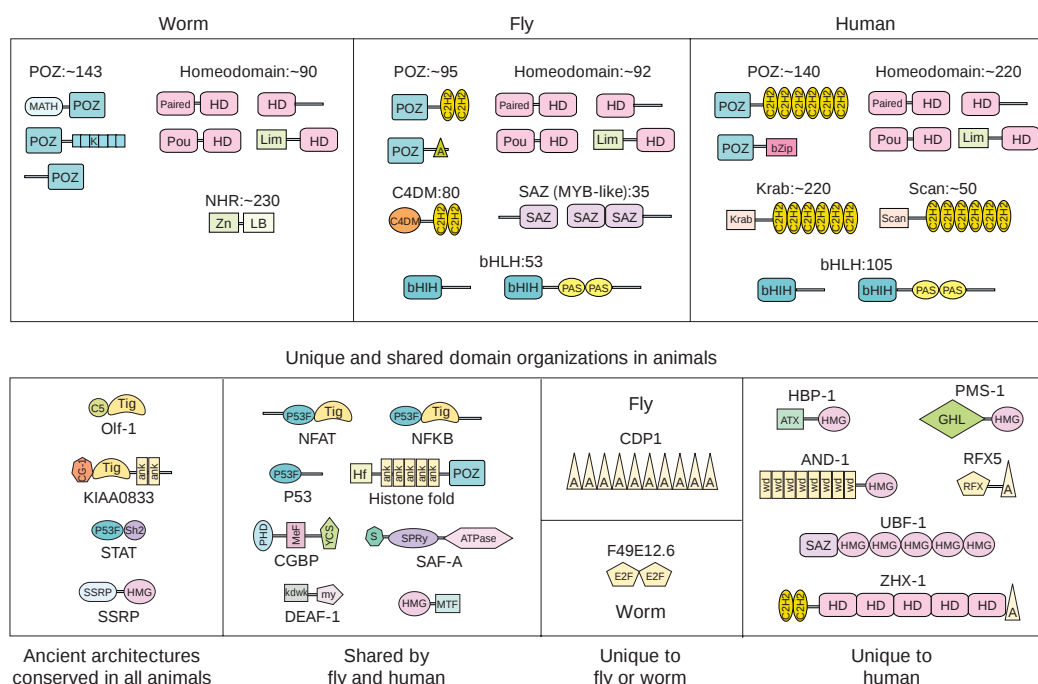


Figure 45 Lineage-specific expansions of domains and architectures of transcription factors. Top, specific families of transcription factors that have been expanded in each of the proteomes. Approximate numbers of domains identified in each of the (nearly) complete proteomes representing the lineages are shown next to the domains, and some of the most common architectures are shown. Some are shared by different animal lineages; others are lineage-specific. Bottom, samples of architectures from transcription factors that are shared by all animals (ancient architectures), shared by fly and human and unique to each lineage. Domains: K, kelch; HD, homeodomain; Zn, zinc-binding domain; LB, ligand-binding domain; C4DM, novel Zn cluster with four cysteines, probably involved in protein–protein interactions (L. Aravind, unpublished); MATH, meprin-associated TRAF

domain; CG-1, novel domain in KIAA0909-like transcription factors (L. Aravind, unpublished); MTF, myelin transcription factor domain; SAZ, specialized Myb-like helix-turn-helix (HTH) domain found in Stonewall, ADF-1 and Zeste (L. Aravind, unpublished); A, AT-hook motif; E2F, winged HTH DNA-binding domain; GHL, gyrase-histidine kinase-MutL ATPase domain; ATX, ATaXin domain; RFX, RFX winged HTH DNA binding domain; My, MYND domain; KDWK, KDWK DNA-binding domain; POZ, Pox zinc finger domain; S, SAP domain; P53F, P53 fold domain; HF, histone fold; ANK, ankyrin repeat; TIG, transcription factor Ig domain; SSRP, structure-specific recognition protein domain; C5, 5-cysteine metal binding domain; C2H2, classic zinc finger domain; WD, WD40 repeats.

The most extreme mechanism is whole-genome duplication (WGD), through a polyploidization event in which a diploid organism becomes tetraploid. Such events are classified as autopolyploidy or allopolyploidy, depending on whether they involve hybridization between members of the same species or different species. Polyploidization is common in the plant kingdom, with many known examples among wild and domesticated crop species. Alfalfa (*Medicago sativa*) is a naturally occurring autotetraploid³⁶⁴, and *Nicotiana tabacum*, some species of cotton (*Gossypium*) and several of the common brassicas are allotetraploids containing pairs of 'homeologous' chromosome pairs.

In principle, WGD provides the raw material for great bursts of innovation by allowing the duplication and divergence of entire pathways. Ohno³⁶⁵ suggested that WGD has played a key role in evolution. There is evidence for an ancient WGD event in the ancestry of yeast and several independent such events in the ancestry of mustard weed^{366–369}. Such ancient WGD events can be hard to detect because only a minority of the duplicated loci may be retained, with the result that the genes in duplicated segments cannot be aligned in a one-to-one correspondence but rather require many gaps. In addition, duplicated segments may be subsequently rearranged. For example, the ancient duplication in the yeast genome appears to have been followed by loss of more than 90% of the newly duplicated genes³⁶⁶.

One of the most controversial hypotheses about vertebrate

evolution is the proposal that two WGD events occurred early in the vertebrate lineage, around the time of jawed fishes some 500 Myr ago. Some authors^{370–373} have seen support for this theory in the fact that many human genes occur in sets of four homologues—most notably the four extensive HOX gene clusters on chromosomes 2, 7, 12 and 17, whose duplication dates to around the correct time. However, other authors have disputed this interpretation³⁷⁴, suggesting that these cases may reflect unrelated duplications of specific regions rather than successive WGD.

We analysed the draft genome sequence for evidence that might bear on this question. The analysis provides many interesting observations, but no convincing evidence of ancient WGD. We looked for evidence of pairs of chromosomal regions containing many homologous genes. Although we found many pairs containing a few homologous genes, the human genome does not appear to contain any pairs of regions where the density of duplicated genes approaches the densities seen in yeast or mustard weed^{366–369}.

We also examined human proteins in the IPI for which the orthologues among fly or worm proteins occur in the ratios 2:1:1, 3:1:1, 4:1:1 and so on (Fig. 49). The number of such families falls smoothly, with no peak at four and some instances of five or more homologues. Although this does not rule out two rounds of WGD followed by extensive gene loss and some unrelated gene duplication, it provides no support for the theory. More probatively, if two successive rounds of genome duplication occurred, phylogenetic analysis of the proteins having 4:1:1 ratios between human, fly and worm would be expected to show more trees with the topology (A,B)(C,D) for the human sequences than (A,(B,(C,D)))³⁷⁵. However, of 57 sets studied carefully, only 24% of the trees constructed from the 4:1:1 set have the former topology; this is not significantly different from what would be expected under the hypothesis of random sequential duplication of individual loci.

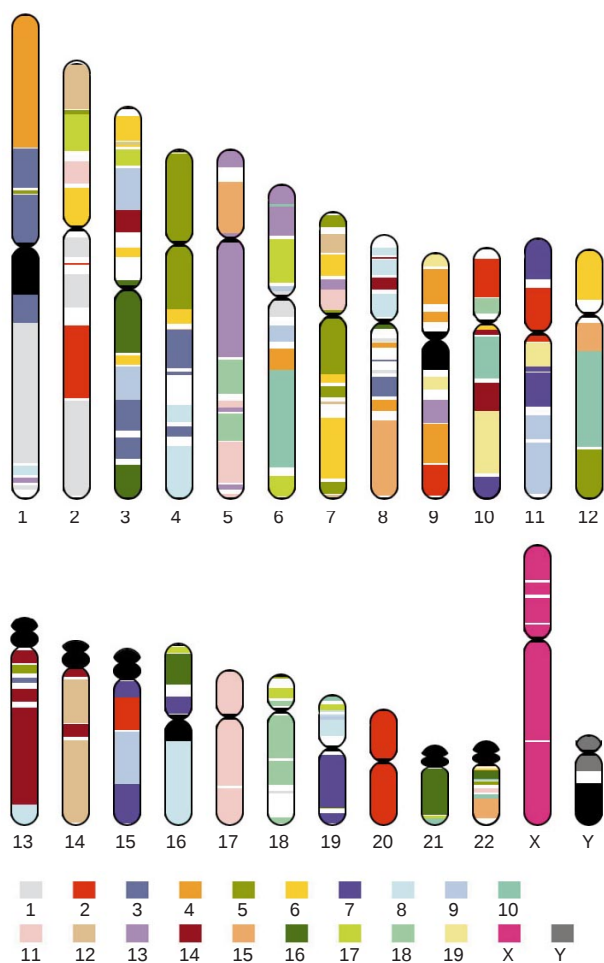


Figure 46 Conserved segments in the human and mouse genome. Human chromosomes, with segments containing at least two genes whose order is conserved in the mouse genome as colour blocks. Each colour corresponds to a particular mouse chromosome. Centromeres, subcentromeric heterochromatin of chromosomes 1, 9 and 16, and the repetitive short arms of 13, 14, 15, 21 and 22 are in black.

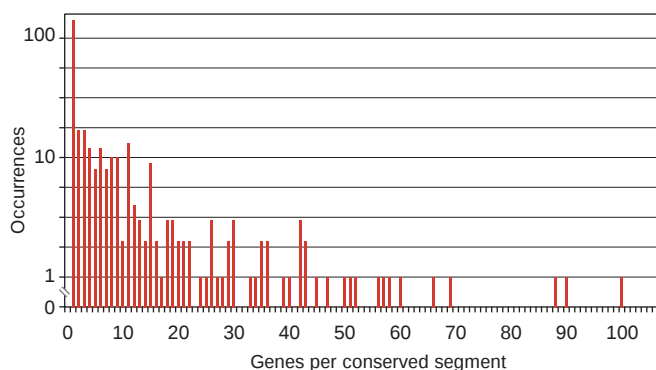


Figure 47 Distribution of number of genes per conserved segment between human and mouse genomes.

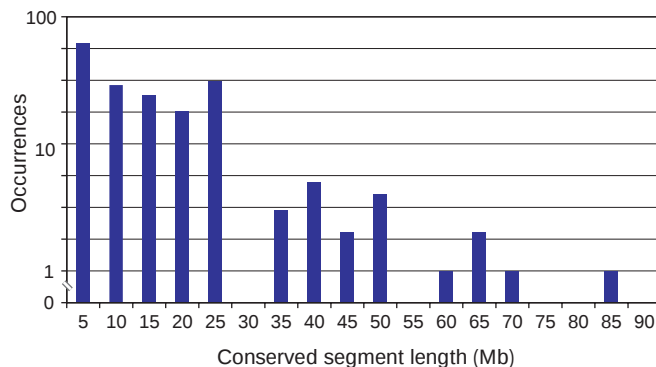


Figure 48 Distribution of lengths (in 5-Mb bins) of conserved segments between human and mouse genomes, omitting singletons.

We also searched for sets of four chromosomes where there are multiple genes with homologues on each of the four. The strongest example was chromosomes 2, 7, 12 and 17, containing the HOX clusters as well as additional genes. These four chromosomes appear to have an excess of quadruplicated genes. The genes are not all clustered in a single region; this may reflect intrachromosomal rearrangement since the duplication of these genes, or it may indicate that they result from several independent events. Of the genes with homologues on chromosomes 2, 12 and 17, many of those missing on chromosome 7 are clustered on chromosome 3, suggesting a translocation. Several additional examples of groups of four chromosomes were found, although they were connected by fewer homologous genes.

Although the analyses are sensitive to the imperfect quality of the gene predictions, our results so far are insufficient to settle whether two rounds of WGD occurred around 500 Myr ago. It may be possible to resolve the issue by systematically estimating the time of each of the many gene duplication events on the basis of sequence divergence, although this is beyond the scope of this report. Another approach to determining whether a widespread duplication occurred at a particular time in vertebrate evolution would be to sequence the genomes of organisms whose lineages diverged from vertebrates at appropriate times, such as amphioxus.

Recent history from human polymorphism

The recent history of genomic segments can be probed by studying the properties of SNPs segregating in the current human population. The sequence information generated in the course of this project has yielded a huge collection of SNPs. These SNPs were extracted in two ways: by comparing overlapping large-insert clones derived from distinct haplotypes (either different individuals or different chromosomes within an individual) and by comparing random reads from whole-genome shotgun libraries derived from multiple individuals. The analysis confirms an average heterozygosity rate in the human population of about 1 in 1,300 bp (ref. 97).

More than 1.42 million SNPs have been assembled into a genome-wide map and are analysed in detail in an accompanying paper⁹⁷. SNP density is also displayed across the genome in Fig. 9. The SNPs have an average spacing of 1.9 kb and 63% of 5-kb intervals contain a SNP. These polymorphisms are of immediate utility for medical genetic studies. Whereas investigators studying a gene previously had to expend considerable effort to discover polymorphisms across the region of interest, the current collection now provides then with about 15 SNPs for gene loci of average size.

The density of SNPs (adjusted for ascertainment—that is, polymorphisms per base screened) varies considerably across the

genome⁹⁷ and sheds light on the unique properties and history of each genomic region. The average heterozygosity at a locus will tend to increase in proportion to the local mutation rate and the 'age' of the locus (which can be defined as the average number of generations since the most recent common ancestor of two randomly chosen copies in the population). For example, positive selection can cause a locus to be unusually 'young' and balancing selection can cause it to be unusually 'old'. An extreme example is the HLA region, in which a high SNP density is observed, reflecting the fact that diverse HLA haplotypes have been maintained for many millions of years by balancing selection and greatly predate the origin of the human species.

SNPs can also be used to study linkage disequilibrium in the human genome³⁷⁶. Linkage disequilibrium refers to the persistence of ancestral haplotypes—that is, genomic segments carrying particular combinations of alleles descended from a common ancestor. It can provide a powerful tool for mapping disease genes^{377,378} and for probing population history^{379–381}. There has been considerably controversy concerning the typical distance over which linkage disequilibrium extends in the human genome^{382–387}. With the collection of SNPs now available, it should be possible to resolve this important issue.

Applications to medicine and biology

In most research papers, the authors can only speculate about future applications of the work. Because the genome sequence has been released on a daily basis over the past four years, however, we can already cite many direct applications. We focus on a handful of applications chosen primarily from medical research.

Disease genes

A key application of human genome research has been the ability to find disease genes of unknown biochemical function by positional cloning³⁸⁸. This method involves mapping the chromosomal region containing the gene by linkage analysis in affected families and then scouring the region to find the gene itself. Positional cloning is powerful, but it has also been extremely tedious. When the approach was first proposed in the early 1980s⁹, a researcher wishing to perform positional cloning had to generate genetic markers to trace inheritance; perform chromosomal walking to obtain genomic DNA covering the region; and analyse a region of around 1 Mb by either direct sequencing or indirect gene identification methods. The first two barriers were eliminated with the development in the mid-1990s of comprehensive genetic and physical maps of the human chromosomes, under the auspices of the Human Genome Project. The remaining barrier, however, has continued to be formidable.

All that is changing with the availability of the human draft genome sequence. The human genomic sequence in public databases allows rapid identification *in silico* of candidate genes, followed by mutation screening of relevant candidates, aided by information on gene structure. For a mendelian disorder, a gene search can now often be carried out in a matter of months with only a modestly sized team.

At least 30 disease genes^{55,389–422} (Table 26) have been positionally cloned in research efforts that depended directly on the publicly available genome sequence. As most of the human sequence has only arrived in the past twelve months, it is likely that many similar discoveries are not yet published. In addition, there are many cases in which the genome sequence played a supporting role, such as providing candidate microsatellite markers for finer genetic linkage analysis.

The genome sequence has also helped to reveal the mechanisms leading to some common chromosomal deletion syndromes. In several instances, recurrent deletions have been found to result from homologous recombination and unequal crossing over between large, nearly identical intrachromosomal duplications. Examples

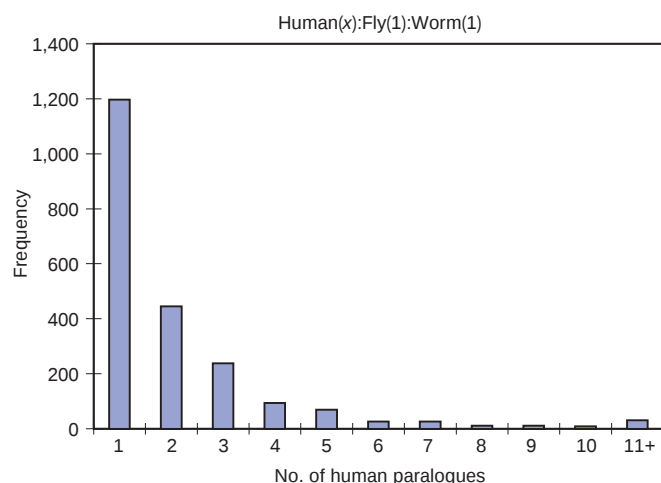


Figure 49 Number of human paralogs of genes having single orthologues in worm and fly.

include the DiGeorge/velocardiofacial syndrome region on chromosome 22 (ref. 238) and the Williams–Beuren syndrome recurrent deletion on chromosome 7 (ref. 239).

The availability of the genome sequence also allows rapid identification of paralogues of disease genes, which is valuable for two reasons. First, mutations in a paralogous gene may give rise to a related genetic disease. A good example, discovered through use of the genome sequence, is achromatopsia (complete colour blindness). The *CNGA3* gene, encoding the α -subunit of the cone photoreceptor cyclic GMP-gated channel, had been shown to harbour mutations in some families with achromatopsia. Computational searching of the genome sequences revealed the paralogous gene encoding the corresponding β -subunit, *CNGB3* (which had not been apparent from EST databases). The *CNGB3* gene was rapidly shown to be the cause of achromatopsia in other families^{407,408}. Another example is provided by the presenilin-1 and presenilin-2 genes, in which mutations can cause early-onset Alzheimer's disease^{423,424}. Second, the paralogue may provide an opportunity for therapeutic intervention, as exemplified by attempts to reactivate the fetally expressed haemoglobin genes in individuals with sickle cell disease or β -thalassaemia, caused by mutations in the β -globin gene⁴²⁵.

We undertook a systematic search for paralogues of 971 known human disease genes with entries in both the Online Mendelian Inheritance in Man (OMIM) database (<http://www.ncbi.nlm.nih.gov/Omim/>) and either the SwissProt or TrEMBL protein databases. We identified 286 potential paralogues (with the requirement of a match of at least 50 amino acids with identity greater than 70% but less than 90% if on the same chromosome, and less than 95% if on a different chromosome). Although this analysis may have identified some pseudogenes, 89% of the matches showed homology over more than one exon in the new target sequence, suggesting that many are functional. This analysis shows the potential for rapid identification of disease gene paralogues *in silico*.

Drug targets

Over the past century, the pharmaceutical industry has largely depended upon a limited set of drug targets to develop new

therapies. A recent compendium^{426,427} lists 483 drug targets as accounting for virtually all drugs on the market. Knowing the complete set of human genes and proteins will greatly expand the search for suitable drug targets. Although only a minority of human genes may be drug targets, it has been predicted that the number will exceed several thousand, and this prospect has led to a massive expansion of genomic research in pharmaceutical research and development. A few examples will illustrate the point.

(1) The neurotransmitter serotonin (5-HT) mediates rapid excitatory responses through ligand-gated channels. The previously identified 5-HT_{3A} receptor gene produces functional receptors, but with a much smaller conductance than observed *in vivo*. Cross-hybridization experiments and analysis of ESTs failed to reveal any other homologues of the known receptor. Recently, however, by searching the human draft genome sequence at low stringency, a putative homologue was identified within a PAC clone from the long arm of chromosome 11 (ref. 428). The homologue was shown to be expressed in the amygdala, caudate and hippocampus, and a full-length cDNA was subsequently obtained. The gene, which codes for a serotonin receptor, was named 5-HT_{3B}. When assembled in a heterodimer with 5-HT_{3A}, it was shown to account for the large-conductance neuronal serotonin channel. Given the central role of the serotonin pathway in mood disorders and schizophrenia, the discovery of a major new therapeutic target is of considerable interest.

(2) The contractile and inflammatory actions of the cysteinyl leukotrienes, formerly known as the slow reacting substance of anaphylaxis (SRS-A), are mediated through specific receptors. The second such receptor, CysLT₂, was identified using the combination of a rat EST and the human genome sequence. This led to the cloning of a gene with 38% amino-acid identity to the only other receptor that had previously been identified⁴²⁹. This new receptor, which shows high-affinity binding to several leukotrienes, maps to a region of chromosome 13 that is linked to atopic asthma. The gene is expressed in airway smooth muscles and in the heart. As the leukotriene pathway has been a significant target for the development of drugs against asthma, the discovery of a new receptor has obvious and important consequences.

(3) Abundant deposition of β -amyloid in senile plaques is the hallmark of Alzheimer's disease. β -Amyloid is generated by proteolytic processing of the amyloid precursor protein (APP). One of the enzymes involved is the β -site APP-cleaving enzyme (BACE), which is a transmembrane aspartyl protease. Computational searching of the public human draft genome sequence recently identified a new sequence homologous to BACE, encoding a protein now named BACE2^{430,431}. BACE2, which has 52% amino-acid sequence identity to BACE, contains two active protease sites and maps to the obligatory Down's syndrome region of chromosome 21, as does APP. This raises the question of whether the extra copies of both BACE2 and APP may contribute to accelerated deposition of β -amyloid in the brains of Down's syndrome patients. The development of antagonists to BACE and BACE2 represents a promising approach to preventing Alzheimer's disease.

Given these examples, we undertook a systematic effort to identify paralogues of the classic drug target proteins in the draft genome sequence. The target list⁴²⁷ was used to identify 603 entries in the SwissProt database with unique accession numbers. These were then searched against the current genome sequence database, using the requirement that a match should have 70–100% identity to at least 50 amino acids. Matches to named proteins were ignored, as we assumed that these represented known homologues.

We found 18 putative novel paralogues (Table 27), including apparent dopamine receptors, purinergic receptors and insulin-like growth factor receptors. In six cases, the novel paralogue matches at least one EST, adding confidence that this search process can identify novel functional genes. For the remaining 12 putative paralogues without an EST match, all have long ORFs and all but

Table 26 Disease genes positionally cloned using the draft genome sequence

Locus	Disorder	Reference(s)
<i>BRCA2</i>	Breast cancer susceptibility	55
<i>AIRE</i>	Autoimmune polyglandular syndrome type 1 (APS1 or APECED)	389
<i>PEX1</i>	Peroxisome biogenesis disorder	390, 391
<i>PDS</i>	Pendred syndrome	392
<i>XLP</i>	X-linked lymphoproliferative disease	393
<i>DFNA5</i>	Nonsyndromic deafness	394
<i>ATP2A2</i>	Darier's disease	395
<i>SEDL</i>	X-linked spondyloepiphyseal dysplasia tarda	396
<i>WISP3</i>	Progressive pseudorheumatoid dysplasia	397
<i>CCM1</i>	Cerebral cavernous malformations	398, 399
<i>COL11A2/DFNA13</i>	Nonsyndromic deafness	400
<i>LGMD 2G</i>	Limb-girdle muscular dystrophy	401
<i>EVC</i>	Ellis-Van Creveld syndrome, Weyer's acrodistal dysostosis	402
<i>ACTN4</i>	Familial focal segmental glomerulosclerosis	403
<i>SCN1A</i>	Generalized epilepsy with febrile seizures plus type 2	404
<i>AASS</i>	Familial hyperlipidaemia	405
<i>NDRG1</i>	Hereditary motor and sensory neuropathy-Lom	406
<i>CNGB3</i>	Total colour-blindness	407, 408
<i>MUL</i>	Mulibrey nanism	409
<i>USH1C</i>	Usher type 1C	410, 411
<i>MYH9</i>	May-Hegglin anomaly	412, 413
<i>PRKAR1A</i>	Carney's complex	414
<i>MYH9</i>	Nonsyndromic hereditary deafness DFNA17	415
<i>SCA10</i>	Spinocerebellar ataxia type 10	416
<i>OPA1</i>	Optic atrophy	417
<i>XLCSNB</i>	X-linked congenital stationary night blindness	418
<i>FGF23</i>	Hypophosphataemic rickets	419
<i>GAN</i>	Giant axonal neuropathy	420
<i>AAAS</i>	Triple-A syndrome	421
<i>HSPG2</i>	Schwartz-Jampel syndrome	422

one show similarity spanning multiple exons separated by introns, so these are not processed pseudogenes. They are likely to represent interesting new candidate drug targets.

Basic biology

Although the examples above reflect medical applications, there are also many similar applications to basic physiology and cell biology. To cite one satisfying example, the publicly available sequence was used to solve a mystery that had vexed investigators for several decades: the molecular basis of bitter taste⁴³². Humans and other animals are polymorphic for response to certain bitter tastes. Recently, investigators mapped this trait in both humans and mice and then searched the relevant region of the human draft genome sequence for G-protein coupled receptors. These studies led, in quick succession, to the discovery of a new family of such proteins, the demonstration that they are expressed almost exclusively in taste buds, and the experimental confirmation that the receptors in cultured cells respond to specific bitter substances^{433–435}.

The next steps

Considerable progress has been made in human sequencing, but much remains to be done to produce a finished sequence. Even more work will be required to extract the full information contained in the sequence. Many of the key next steps are already underway.

Finishing the human sequence

The human sequence will serve as a foundation for biomedical research in the years ahead, and it is thus crucial that the remaining gaps be filled and ambiguities be resolved as quickly as possible. This will involve a three-step program.

The first stage involves producing finished sequence from clones spanning the current physical map, which covers more than 96% of the euchromatic regions of the genome. About 1 Gb of finished sequence is already completed. Almost all of the remaining clones are already sequenced to at least draft coverage, and the rest have been selected for sequencing. All clones are expected to reach 'full shotgun' coverage (8–10-fold redundancy) by about mid-2001 and finished form (99.99% accuracy) not long thereafter, using established and increasingly automated protocols.

The next stage will be to screen additional libraries to close gaps between clone contigs. Directed probing of additional large-insert clone libraries should close many of the remaining gaps. Unclosed gaps will be sized by FISH techniques or other methods. Two chromosomes, 22 and 21, have already been assembled in this 'essentially complete' form in this manner^{93,94}, and chromosomes 20, Y, 19, 14 and 7 are likely to reach this status in the next few

months. All chromosomes should be essentially completed by 2003, if not sooner.

Finally, techniques must be developed to close recalcitrant gaps. Several hundred such gaps in the euchromatic sequence will probably remain in the genome after exhaustive screening of existing large-insert libraries. New methodologies will be needed to recover sequence from these segments, and to define biological reasons for their lack of representation in standard libraries. Ideally, it would be desirable to obtain complete sequence from all heterochromatic regions, such as centromeres and ribosomal gene clusters, although most of this sequence will consist of highly polymorphic tandem repeats containing few protein-coding genes.

Developing the IGI and IPI

The draft genome sequence has provided an initial look at the human gene content, but many ambiguities remain. A high priority will be to refine the IGI and IPI to the point where they accurately reflect every gene and every alternatively spliced form. Several steps are needed to reach this ambitious goal.

Finishing the human sequence will assist in this effort, but the experiences gained on chromosomes 21 and 22 show that sequence alone is not enough to allow complete gene identification. One powerful approach is cross-species sequence comparison with related organisms at suitable evolutionary distances. The sequence coverage from the pufferfish *T. nigroviridis* has already proven valuable in identifying potential exons²⁹²; this work is expected to continue from its current state of onefold coverage to reach at least fivefold coverage later this year. The genome sequence of the laboratory mouse will provide a particularly powerful tool for exon identification, as sequence similarity is expected to identify 95–97% of the exons, as well as a significant number of regulatory domains^{436–438}. A public-private consortium is speeding this effort, by producing freely accessible whole-genome shotgun coverage that can be readily used for cross-species comparison⁴³⁹. More than onefold coverage from the C57BL/6J strain has already been completed and threefold is expected within the next few months. In the slightly longer term, a program is under way to produce a finished sequence of the laboratory mouse.

Another important step is to obtain a comprehensive collection of full-length human cDNAs, both as sequences and as actual clones. The Mammalian Gene Collection project has been underway for a year¹⁸ and expects to produce 10,000–15,000 human full-length cDNAs over the coming year, which will be available without restrictions on use. The Genome Exploration Group of the RIKEN Genomic Sciences Center is similarly developing a collection of cDNA clones from mouse³⁰⁹, which is a valuable complement

Table 27 New paralogues of common drug targets identified by searching the draft human genome sequence

Drug target	HGM symbol	SwissProt accession	Chromosome	Novel match IGI number	Chromosome containing paralogue	Per cent identity	dbEST GenBank accession
Aquaporin 7	AQP7	O14520	9	IGL_M1_ctg15869_11	2	92.3	AW593324
Arachidonate 12-lipoxygenase	ALOX12	P18054	17	IGL_M1_ctg17216_23	17	70.1	
Calcitonin	CALCA	P01258	11	IGL_M1_ctg14138_20	12	93.6	
Calcium channel, voltage-dependent, γ -subunit	CACNG2	Q9Y698	22	IGL_M1_ctg17137_10	19	70.7	
DNA polymerase- δ , small subunit	POLD2	P49005	7	IGL_M1_ctg12903_29	5	86.8	
Dopamine receptor, D1- α	DRD1	P21728	5	IGL_M1_ctg25203_33	16	70.7	
Dopamine receptor, D1- β	DRD5	P21918	4	IGL_M1_ctg17190_14	1	88.0	AI148329
Eukaryotic translation elongation factor, 1 δ	EEF1D	P29692	2	IGL_M1_ctg16401_37	17	77.6	BE719683
FKBP, tacrolimus binding protein, FK506 binding protein	FKBP1B	Q16645	2	IGL_M1_ctg14291_56	6	79.4	
Glutamic acid decarboxylase	GAD1	Q99259	2	IGL_M1_ctg12341_103	18	70.5	
Glycine receptor, α 1	GLRA1	P23415	5	IGL_M1_ctg16547_14	X	85.5	
Heparan N-deacetylase/N-sulphotransferase	NDST1	P52848	5	IGL_M1_ctg13263_18	4	81.5	
Insulin-like growth factor-1 receptor	IGF1R	P08069	15	IGL_M1_ctg18444_3	19	71.8	
Na,K-ATPase, α -subunit	ATP1A1	P05023	1	IGL_M1_ctg14877_54	1	83	
Purinergic receptor 7 (P2X), ligand-gated ion channel	P2RX7	Q99572	12	IGL_M1_ctg15140_15	12	80.3	H84353
Tubulin, ϵ -chain	TUBE*	Q9UJ70	6	IGL_M1_ctg13826_4	5	78.5	AA970498
Tubulin, χ -chain	TUBG1	P23258	17	IGL_M1_ctg12599_5	7	84.0	
Voltage-gated potassium channel, KV3.3	KCNC3	Q14003	19	IGL_M1_ctg13492_5	12	80.1	H49142

* HGM symbol unknown.

because of the availability of tissues from all developmental time points. A challenge will be to define the gene-specific patterns of alternative splicing, which may affect half of human genes. Existing collections of ESTs and cDNAs may allow identification of the most abundant of these isoforms, but systematic exploration of this problem may require exhaustive analysis of cDNA libraries from multiple tissues or perhaps high-throughput reverse transcription–PCR studies. Deep understanding of gene function will probably require knowledge of the structure, tissue distribution and abundance of these alternative forms.

Large-scale identification of regulatory regions

The one-dimensional script of the human genome, shared by essentially all cells in all tissues, contains sufficient information to provide for differentiation of hundreds of different cell types, and the ability to respond to a vast array of internal and external influences. Much of this plasticity results from the carefully orchestrated symphony of transcriptional regulation. Although much has been learned about the *cis*-acting regulatory motifs of some specific genes, the regulatory signals for most genes remain uncharacterized. Comparative genomics of multiple vertebrates offers the best hope for large-scale identification of such regulatory sites⁴⁴⁰. Previous studies of sequence alignment of regulatory domains of orthologous genes in multiple species has shown a remarkable correlation between sequence conservation, dubbed ‘phylogenetic footprints’⁴⁴¹, and the presence of binding motifs for transcription factors. This approach could be particularly powerful if combined with expression array technologies that identify cohorts of genes that are coordinately regulated, implicating a common set of *cis*-acting regulatory sequences^{442–445}. It will also be of considerable interest to study epigenetic modifications such as cytosine methylation on a genome-wide scale, and to determine their biological consequences^{446,447}. Towards this end, a pilot Human Epigenome Project has been launched^{448,449}.

Sequencing of additional large genomes

More generally, comparative genomics allows biologists to peruse evolution’s laboratory notebook—to identify conserved functional features and recognize new innovations in specific lineages. Determination of the genome sequence of many organisms is very desirable. Already, projects are underway to sequence the genomes of the mouse, rat, zebrafish and the pufferfishes *T. nigroviridis* and *Takifugu rubripes*. Plans are also under consideration for sequencing additional primates and other organisms that will help define key developments along the vertebrate and nonvertebrate lineages.

To realize the full promise of comparative genomics, however, it needs to become simple and inexpensive to sequence the genome of any organism. Sequencing costs have dropped 100-fold over the last 10 years, corresponding to a roughly twofold decrease every 18 months. This rate is similar to ‘Moore’s law’ concerning improvements in semiconductor manufacture. In both sequencing and semiconductors, such improvement does not happen automatically, but requires aggressive technological innovation fuelled by major investment. Improvements are needed to move current dideoxy sequencing to smaller volumes and more rapid sequencing times, based upon advances such as microchannel technology. More revolutionary methods, such as mass spectrometry, single-molecule sequencing and nanopore approaches⁷⁶, have not yet been fully developed, but hold great promise and deserve strong encouragement.

Completing the catalogue of human variation

The human draft genome sequence has already allowed the identification of more than 1.4 million SNPs, comprising a substantial proportion of all common human variation. This program should be extended to obtain a nearly complete catalogue of common variants and to identify the common ancestral haplotypes present in the population. In principle, these genetic tools should make it possible to perform association studies and linkage disequilibrium studies³⁷⁶ to identify the genes that confer even relatively modest risk

for common diseases. Launching such an intense era of human molecular epidemiology will also require major advances in the cost efficiency of genotyping technology, in the collection of carefully phenotyped patient cohorts and in statistical methods for relating large-scale SNP data to disease phenotype.

From sequence to function

The scientific program outlined above focuses on how the genome sequence can be mined for biological information. In addition, the sequence will serve as a foundation for a broad range of functional genomic tools to help biologists to probe function in a more systematic manner. These will need to include improved techniques and databases for the global analysis of: RNA and protein expression, protein localization, protein–protein interactions and chemical inhibition of pathways. New computational techniques will be needed to use such information to model cellular circuitry. A full discussion of these important directions is beyond the scope of this paper.

Concluding thoughts

The Human Genome Project is but the latest increment in a remarkable scientific program whose origins stretch back a hundred years to the rediscovery of Mendel’s laws and whose end is nowhere in sight. In a sense, it provides a capstone for efforts in the past century to discover genetic information and a foundation for efforts in the coming century to understand it.

We find it humbling to gaze upon the human sequence now coming into focus. In principle, the string of genetic bits holds long-sought secrets of human development, physiology and medicine. In practice, our ability to transform such information into understanding remains woefully inadequate. This paper simply records some initial observations and attempts to frame issues for future study. Fulfilling the true promise of the Human Genome Project will be the work of tens of thousands of scientists around the world, in both academia and industry. It is for this reason that our highest priority has been to ensure that genome data are available rapidly, freely and without restriction.

The scientific work will have profound long-term consequences for medicine, leading to the elucidation of the underlying molecular mechanisms of disease and thereby facilitating the design in many cases of rational diagnostics and therapeutics targeted at those mechanisms. But the science is only part of the challenge. We must also involve society at large in the work ahead. We must set realistic expectations that the most important benefits will not be reaped overnight. Moreover, understanding and wisdom will be required to ensure that these benefits are implemented broadly and equitably. To that end, serious attention must be paid to the many ethical, legal and social implications (ELSI) raised by the accelerated pace of genetic discovery. This paper has focused on the scientific achievements of the human genome sequencing efforts. This is not the place to engage in a lengthy discussion of the ELSI issues, which have also been a major research focus of the Human Genome Project, but these issues are of comparable importance and could appropriately fill a paper of equal length.

Finally, it has not escaped our notice that the more we learn about the human genome, the more there is to explore.

“We shall not cease from exploration. And the end of all our exploring will be to arrive where we started, and know the place for the first time.”—T. S. Eliot⁴⁵⁰ □

Received 7 December 2000; accepted 9 January 2001.

1. Correns, C. Untersuchungen über die Xenien bei *Zea mays*. *Berichte der Deutsche Botanische Gesellschaft* **17**, 410–418 (1899).
2. De Vries, H. Sur la loi de disjonction des hybrides. *Comptes Rendue Hebdomadaires, Acad. Sci. Paris* **130**, 845–847 (1900).
3. von Tschermack, E. Über Künstliche Kreuzung bei *Pisum sativum*. *Berichte der Deutsche Botanische Gesellschaft* **18**, 232–239. (1900).
4. Sanger, F. *et al.* Nucleotide sequence of bacteriophage Φ X174 DNA. *Nature* **265**, 687–695 (1977).
5. Sanger, F. *et al.* The nucleotide sequence of bacteriophage Φ X174. *J Mol Biol* **125**, 225–246 (1978).

6. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide-sequence of bacteriophage Lambda DNA. *J. Mol. Biol.* **162**, 729–773 (1982).
7. Fiers, W. *et al.* Complete nucleotide sequence of SV40 DNA. *Nature* **273**, 113–120 (1978).
8. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465 (1981).
9. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
10. Olson, M. V. *et al.* Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl Acad. Sci. USA* **83**, 7826–7830 (1986).
11. Coulson, A., Sulston, J., Brenner, S. & Karn, J. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **83**, 7821–7825 (1986).
12. Putney, S. D., Herlihy, W. C. & Schimmel, P. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302**, 718–721 (1983).
13. Milner, R. J. & Sutcliffe, J. G. Gene expression in rat brain. *Nucleic Acids Res.* **11**, 5497–5520 (1983).
14. Adams, M. D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651–1656 (1991).
15. Adams, M. D. *et al.* Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377**, 3–174 (1995).
16. Okubo, K. *et al.* Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.* **2**, 173–179 (1992).
17. Hillier, L. D. *et al.* Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6**, 807–828 (1996).
18. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. The mammalian gene collection. *Science* **286**, 455–457 (1999).
19. Berry, R. *et al.* Gene-based sequence-tagged-sites (STSs) as the basis for a human gene map. *Nature Genet.* **10**, 415–423 (1995).
20. Houlgate, R. *et al.* The Genexpress Index: a resource for gene discovery and the genic map of the human genome. *Genome Res.* **5**, 272–304 (1995).
21. Sinheimer, R. L. The Santa Cruz Workshop—May 1985. *Genomics* **5**, 954–956 (1989).
22. Palca, J. Human genome—Department of Energy on the map. *Nature* **321**, 371 (1986).
23. National Research Council *Mapping and Sequencing the Human Genome* (National Academy Press, Washington DC, 1988).
24. Bishop, J. E. & Waldholz, M. *Genome* (Simon and Schuster, New York, 1990).
25. Keyes, D. J. & Hood, L. (eds) *The Code of Codes: Scientific and Social Issues in the Human Genome Project* (Harvard Univ. Press, Cambridge, Massachusetts, 1992).
26. Cook-Deegan, R. *The Gene Wars: Science, Politics, and the Human Genome* (W. W. Norton & Co., New York, London, 1994).
27. Donis-Keller, H. *et al.* A genetic linkage map of the human genome. *Cell* **51**, 319–337 (1987).
28. Gyapay, G. *et al.* The 1993–94 Genethon human genetic linkage map. *Nature Genet.* **7**, 246–339 (1994).
29. Hudson, T. J. *et al.* An STS-based map of the human genome. *Science* **270**, 1945–1954 (1995).
30. Dietrich, W. F. *et al.* A comprehensive genetic map of the mouse genome. *Nature* **380**, 149–152 (1996).
31. Nusbaum, C. *et al.* A YAC-based physical map of the mouse genome. *Nature Genet.* **22**, 388–393 (1999).
32. Oliver, S. G. *et al.* The complete DNA sequence of yeast chromosome III. *Nature* **357**, 38–46 (1992).
33. Wilson, R. *et al.* 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature* **368**, 32–38 (1994).
34. Chen, E. Y. *et al.* The human growth hormone locus: nucleotide sequence, biology, and evolution. *Genomics* **4**, 479–497 (1989).
35. McCombie, W. R. *et al.* Expressed genes, Alu repeats and polymorphisms in cosmids sequenced from chromosome 4p16.3. *Nature Genet.* **1**, 348–353 (1992).
36. Martin-Gallardo, A. *et al.* Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3. *Nature Genet.* **1**, 34–39 (1992).
37. Edwards, A. *et al.* Automated DNA sequencing of the human HPRT locus. *Genomics* **6**, 593–608 (1990).
38. Marshall, E. A strategy for sequencing the genome 5 years early. *Science* **267**, 783–784 (1995).
39. Project to sequence human genome moves on to the starting blocks. *Nature* **375**, 93–94 (1995).
40. Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA* **89**, 8794–8797 (1992).
41. Burke, D. T., Carle, G. F. & Olson, M. V. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236**, 806–812 (1987).
42. Marshall, E. A second private genome project. *Science* **281**, 1121 (1998).
43. Marshall, E. NIH to produce a ‘working draft’ of the genome by 2001. *Science* **281**, 1774–1775 (1998).
44. Pennisi, E. Academic sequencers challenge Celera in a sprint to the finish. *Science* **283**, 1822–1823 (1999).
45. Bouck, J., Miller, W., Gorrell, J. H., Muzny, D. & Gibbs, R. A. Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Res.* **8**, 1074–1084 (1998).
46. Collins, F. S. *et al.* New goals for the U. S. Human Genome Project: 1998–2003. *Science* **282**, 682–689 (1998).
47. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 441–448 (1975).
48. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA* **74**, 560–564 (1977).
49. Anderson, S. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.* **9**, 3015–3027 (1981).
50. Gardner, R. C. *et al.* The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Res.* **9**, 2871–2888 (1981).
51. Deininger, P. L. Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. *Anal. Biochem.* **129**, 216–223 (1983).
52. Chisoe, S. L. *et al.* Sequence and analysis of the human ABL gene, the BCR gene, and regions involved in the Philadelphia chromosomal translocation. *Genomics* **27**, 67–82 (1995).
53. Rowen, L., Koop, B. F. & Hood, L. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science* **272**, 1755–1762 (1996).
54. Koop, B. F. *et al.* Organization, structure, and function of 95 kb of DNA spanning the murine T-cell receptor C alpha/C delta region. *Genomics* **13**, 1209–1230 (1992).
55. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**, 789–792 (1995).
56. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
57. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).
58. Weber, J. L. & Myers, E. W. Human whole-genome shotgun sequencing. *Genome Res.* **7**, 401–409 (1997).
59. Green, P. Against a whole-genome shotgun. *Genome Res.* **7**, 410–417 (1997).
60. Venter, J. C. *et al.* Shotgun sequencing of the human genome. *Science* **280**, 1540–1542 (1998).
61. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
62. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).
63. Ju, J. Y., Ruan, C. C., Fuller, C. W., Glazer, A. N. & Mathies, R. A. Fluorescence energy-transfer dye-labeled primers for DNA sequencing and analysis. *Proc. Natl Acad. Sci. USA* **92**, 4347–4351 (1995).
64. Lee, L. G. *et al.* New energy transfer dyes for DNA sequencing. *Nucleic Acids Res.* **25**, 2816–2822 (1997).
65. Rosenblum, B. B. *et al.* New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Res.* **25**, 4500–4504 (1997).
66. Metzker, M. L., Lu, J. & Gibbs, R. A. Electrophoretically uniform fluorescent dyes for automated DNA sequencing. *Science* **271**, 1420–1422 (1996).
67. Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336–341 (1987).
68. Reeve, M. A. & Fuller, C. W. A novel thermostable polymerase for DNA sequencing. *Nature* **376**, 796–797 (1995).
69. Tabor, S. & Richardson, C. C. Selective inactivation of the exonuclease activity of bacteriophage T7 DNA polymerase by in vitro mutagenesis. *J. Biol. Chem.* **264**, 6447–6458 (1989).
70. Tabor, S. & Richardson, C. C. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase—effect of pyrophosphorylation and metal ions. *J. Biol. Chem.* **265**, 8322–8328 (1990).
71. Murray, V. Improved double-stranded DNA sequencing using the linear polymerase chain reaction. *Nucleic Acids Res.* **17**, 8889 (1989).
72. Guttman, A., Cohen, A. S., Heiger, D. N. & Karger, B. L. Analytical and micropreparative ultrahigh resolution of oligonucleotides by polyacrylamide-gel high-performance capillary electrophoresis. *Anal. Chem.* **62**, 137–141 (1990).
73. Luckey, J. A. *et al.* High-speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res.* **18**, 4417–4421 (1990).
74. Swerdlow, H., Wu, S., Harke, H. & Dovichi, N. J. Capillary gel-electrophoresis for DNA sequencing—laser-induced fluorescence detection with the sheath flow cuvette. *J. Chromatogr.* **516**, 61–67 (1990).
75. Meldrum, D. Automation for genomics, part one: preparation for sequencing. *Genome Res.* **10**, 1081–1092 (2000).
76. Meldrum, D. Automation for genomics, part two: sequencers, microarrays, and future trends. *Genome Res.* **10**, 1288–1303 (2000).
77. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
78. Ewing, B., Hillier, L., Wendt, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
79. Bentley, D. R. Genomic sequence information should be released immediately and freely in the public domain. *Science* **274**, 533–534 (1996).
80. Guyer, M. Statement on the rapid release of genomic DNA sequence. *Genome Res.* **8**, 413 (1998).
81. Dietrich, W. *et al.* A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics* **131**, 423–447 (1992).
82. Kim, U. J. *et al.* Construction and characterization of a human bacterial artificial chromosome library. *Genomics* **34**, 213–218 (1996).
83. Osoegawa, K. *et al.* Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res.* **10**, 116–128 (2000).
84. Marra, M. A. *et al.* High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7**, 1072–1084 (1997).
85. Marra, M. A. *et al.* A map for sequence analysis of the *Arabidopsis thaliana* genome. *Nature Genet.* **22**, 265–270 (1999).
86. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
87. Zhao, S. *et al.* Human BAC ends quality assessment and sequence analyses. *Genomics* **63**, 321–332 (2000).
88. Mahairas, G. G. *et al.* Sequence-tagged connectors: A sequence approach to mapping and scanning the human genome. *Proc. Natl Acad. Sci. USA* **96**, 9739–9744 (1999).
89. Tilford, C. A. *et al.* A physical map of the human Y chromosome. *Nature* **409**, 943–945 (2001).
90. Bentley, D. R. *et al.* The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409**, 942–943 (2001).
91. Montgomery, K. T. *et al.* A high-resolution map of human chromosome 12. *Nature* **409**, 945–946 (2001).
92. Brails, T. *et al.* A physical map of human chromosome 14. *Nature* **409**, 947–948 (2001).
93. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
94. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
95. Cox, D. *et al.* Radiation hybrid map of the human genome. *Science* (in the press).
96. Osoegawa, K. *et al.* An improved approach for construction of bacterial artificial chromosome libraries. *Genomics* **52**, 1–8 (1998).
97. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
98. Collins, F. S., Brooks, L. D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**, 1229–1231 (1998).
99. Stewart, E. A. *et al.* An STS-based radiation hybrid map of the human genome. *Genome Res.* **7**, 422–433 (1997).
100. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
101. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites.

- Nature* **380**, 152–154 (1996).
102. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
103. The BAC Resource Consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**, 953–958 (2001).
104. Kent, W. J. & Haussler, D. GigAssembler: an algorithm for the initial assembly of the human working draft. Technical Report UCSC-CRL-00-17 (Univ. California at Santa Cruz, Santa Cruz, California, 2001).
105. Morton, N. E. Parameters of the human genome. *Proc. Natl Acad. Sci. USA* **88**, 7474–7476 (1991).
106. Podugolnikova, O. A. & Blumina, M. G. Heterochromatic regions on chromosomes 1, 9, 16, and Y in children with some disturbances occurring during embryo development. *Hum. Genet.* **63**, 183–188 (1983).
107. Lundgren, R., Berger, R. & Kristoffersson, U. Constitutive heterochromatin C-band polymorphism in prostatic cancer. *Cancer Genet. Cytogenet.* **51**, 57–62 (1991).
108. Lee, C., Wevrick, R., Fisher, R. B., Ferguson-Smith, M. A. & Lin, C. C. Human centromeric DNAs. *Hum. Genet.* **100**, 291–304 (1997).
109. Riethman, H. C. *et al.* Integration of telomere sequences with the draft human genome sequence. *Nature* **409**, 953–958 (2001).
110. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).
111. Wolfsberg, T. G., McEntyre, J. & Schuler, G. D. Guide to the draft human genome. *Nature* **409**, 824–826 (2001).
112. Hurst, L. D. & Eyre-Walker, A. Evolutionary genomics: reading the bands. *Bioessays* **22**, 105–107 (2000).
113. Saccone, S. *et al.* Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl Acad. Sci. USA* **90**, 11929–11933 (1993).
114. Zoubak, S., Clay, O. & Bernardi, G. The gene distribution of the human genome. *Gene* **174**, 95–102 (1996).
115. Gardiner, K. Base composition and gene distribution: critical patterns in mammalian genome organization. *Trends Genet.* **12**, 519–524 (1996).
116. Duret, L., Mouchiroud, D. & Gautier, C. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.* **40**, 308–317 (1995).
117. Saccone, S., De Sario, A., Della Valle, G. & Bernardi, G. The highest gene concentrations in the human genome are in telomeric bands of metaphase chromosomes. *Proc. Natl Acad. Sci. USA* **89**, 4913–4917 (1992).
118. Bernardi, G. *et al.* The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953–958 (1985).
119. Bernardi, G. Isochores and the evolutionary genomics of vertebrates. *Gene* **241**, 3–17 (2000).
120. Fickett, J. W., Torney, D. C. & Wolf, D. R. Base compositional structure of genomes. *Genomics* **13**, 1056–1064 (1992).
121. Churchill, G. A. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51**, 79–94 (1989).
122. Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**, 91–99 (1985).
123. Bird, A. P. CpG islands as gene markers in the vertebrate nucleus. *Trends Genet.* **3**, 342–347 (1987).
124. Chan, M. F., Liang, G. & Jones, P. A. Relationship between transcription and DNA methylation. *Curr. Top. Microbiol. Immunol.* **249**, 75–86 (2000).
125. Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187**, 226–232 (1975).
126. Larsen, F., Gundersen, G., Lopez, R. & Prydz, H. CpG islands as gene markers in the human genome. *Genomics* **13**, 1095–1107 (1992).
127. Tazi, J. & Bird, A. Alternative chromatin structure at CpG islands. *Cell* **60**, 909–920 (1990).
128. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
129. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90**, 11995–11999 (1993).
130. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25**, 232–234 (2000).
131. Yu, A. Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953 (2001).
132. Kaback, D. B., Guacci, V., Barber, D. & Mahon, J. W. Chromosome size-dependent control of meiotic recombination. *Science* **256**, 228–232 (1992).
133. Riles, L. *et al.* Physical maps of the 6 smallest chromosomes of *Saccharomyces cerevisiae* at a resolution of 2.6-kilobase pairs. *Genetics* **134**, 81–150 (1993).
134. Lynn, A. *et al.* Patterns of meiotic recombination on the long arm of human chromosome 21. *Genome Res.* **10**, 1319–1332 (2000).
135. Laurie, D. A. & Hulten, M. A. Further studies on bivalent chiasma frequency in human males with normal karyotypes. *Ann. Hum. Genet.* **49**, 189–201 (1985).
136. Roeder, G. S. Meiotic chromosomes: it takes two to tango. *Genes Dev.* **11**, 2600–2621 (1997).
137. Wu, T.-C. & Lichten, M. Meiosis-induced double-strand break sites determined by yeast chromatin structure. *Science* **263**, 515–518 (1994).
138. Gerton, J. L. *et al.* Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **97**, 11383–11390 (2000).
139. Li, W.-H. *Molecular Evolution* (Sinauer, Sunderland, Massachusetts, 1997).
140. Gregory, T. R. & Hebert, P. D. The modulation of DNA content: proximate causes and ultimate consequences. *Genome Res.* **9**, 317–324 (1999).
141. Hartl, D. L. Molecular melodies in high and low C. *Nature Rev. Genet.* **1**, 145–149 (2000).
142. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**, 657–663 (1999).
143. Prak, E. L. & Haig, H. K. Jr Mobile elements and the human genome. *Nature Rev. Genet.* **1**, 134–144 (2000).
144. Okada, N., Hamada, M., Ogiwara, I. & Ohshima, K. SINEs and LINEs share common 3' sequences: a review. *Gene* **205**, 229–243 (1997).
145. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24**, 363–367 (2000).
146. Wei, W. *et al.* Human L1 retrotransposition: cis-preference vs. trans-complementation. *Mol. Cell. Biol.* **21**, 1429–1439 (2001).
147. Malik, H. S., Henikoff, S. & Eickbush, T. H. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res.* **10**, 1307–1318 (2000).
148. Smit, A. F. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6**, 743–748 (1996).
149. Clark, J. B. & Tidwell, M. G. A phylogenetic perspective on P transposable element evolution in *Drosophila*. *Proc. Natl Acad. Sci. USA* **94**, 11428–11433 (1997).
150. Haring, E., Hagemann, S. & Pinsker, W. Ancient and recent horizontal invasions of *Drosophilids* by P elements. *J. Mol. Evol.* **51**, 577–586 (2000).
151. Koga, A. *et al.* Evidence for recent invasion of the medaka fish genome by the Tol2 transposable element. *Genetics* **155**, 273–281 (2000).
152. Robertson, H. M. & Lampe, D. J. Recent horizontal transfer of a mariner transposable element among and between Diptera and Neuroptera. *Mol. Biol. Evol.* **12**, 850–862 (1995).
153. Simmons, G. M. Horizontal transfer of hobo transposable elements within the *Drosophila melanogaster* species complex: evidence from DNA sequencing. *Mol. Biol. Evol.* **9**, 1050–1060 (1992).
154. Malik, H. S., Burke, W. D. & Eickbush, T. H. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16**, 793–805 (1999).
155. Kordis, D. & Gubensek, F. Bov-B long interspersed repeated DNA (LINE) sequences are present in *Vipera ammodytes* phospholipase A2 genes and in genomes of Viperidae snakes. *Eur. J. Biochem.* **246**, 772–779 (1997).
156. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
157. Sarich, V. M. & Wilson, A. C. Generation time and genome evolution in primates. *Science* **179**, 1144–1147 (1973).
158. Smit, A. F., Toth, G., Riggs, A. D., & Jurka, J. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* **246**, 401–417 (1995).
159. Lim, J. K. & Simmons, M. J. Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. *Bioessays* **16**, 269–275 (1994).
160. Caceres, M., Ranz, J. M., Barbado, A., Long, M. & Ruiz, A. Generation of a widespread *Drosophila* inversion by a transposable element. *Science* **285**, 415–418 (1999).
161. Gray, Y. H. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet.* **16**, 461–468 (2000).
162. Zhang, J. & Peterson, T. Genome rearrangements by nonlinear transposons in maize. *Genetics* **153**, 1403–1410 (1999).
163. Smit, A. F. Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res.* **21**, 1863–1872 (1993).
164. Cordonnier, A., Casella, J. F. & Heidmann, T. Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *J. Virol.* **69**, 5890–5897 (1995).
165. Medstrand, P. & Mager, D. L. Human-specific integrations of the HERV-K endogenous retrovirus family. *J. Virol.* **72**, 9782–9787 (1998).
166. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
167. Petrov, D. A., Lozovskaya, E. R. & Hartl, D. L. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**, 346–349 (1996).
168. Li, W. H., Ellsworth, D. L., Krushkal, J., Chang, B. H. & Hewett-Emmett, D. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phylogenet. Evol.* **5**, 182–187 (1996).
169. Goodman, M. *et al.* Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Mol. Phylogenet. Evol.* **9**, 585–598 (1998).
170. Kazanian, H. H. Jr & Moran, J. V. The impact of L1 retrotransposons on the human genome. *Nature Genet.* **19**, 19–24 (1998).
171. Malik, H. S. & Eickbush, T. H. NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from *Caenorhabditis elegans*. *Genetics* **154**, 193–203 (2000).
172. Casavant, N. C. *et al.* The end of the LINE?: lack of recent L1 activity in a group of South American rodents. *Genetics* **154**, 1809–1817 (2000).
173. Meunier-Rotival, M., Soriano, P., Cuny, G., Strauss, F. & Bernardi, G. Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. *Proc. Natl Acad. Sci. USA* **79**, 355–359 (1982).
174. Soriano, P., Meunier-Rotival, M. & Bernardi, G. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **80**, 1816–1820 (1983).
175. Goldman, M. A., Holmquist, G. P., Gray, M. C., Caston, L. A. & Nag, A. Replication timing of genes and middle repetitive sequences. *Science* **224**, 686–692 (1984).
176. Manueldis, L. & Ward, D. C. Chromosomal and nuclear distribution of the *HindIII* 1.9-kb human DNA repeat segment. *Chromosoma* **91**, 28–38 (1984).
177. Feng, Q., Moran, J. V., Kazanian, H. H. Jr & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905–916 (1996).
178. Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retrotransposons. *Proc. Natl Acad. Sci. USA* **94**, 1872–1877 (1997).
179. Arcot, S. S. *et al.* High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. *J. Mol. Biol.* **281**, 843–856 (1998).
180. Schmid, C. W. Does SINE evolution preclude Alu function? *Nucleic Acids Res.* **26**, 4541–4550 (1998).
181. Chu, W. M., Ballard, R., Carpick, B. W., Williams, B. R. & Schmid, C. W. Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR. *Mol. Cell. Biol.* **18**, 58–68 (1998).
182. Li, T., Spearow, J., Rubin, C. M. & Schmid, C. W. Physiological stresses increase mouse short interspersed element (SINE) RNA expression in vivo. *Gene* **239**, 367–372 (1999).
183. Liu, W. M., Chu, W. M., Choudary, P. V. & Schmid, C. W. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res.* **23**, 1758–1765 (1995).
184. Filipski, J. Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Lett.* **217**, 184–186 (1987).
185. Sueoka, N. Directional mutation pressure and neutral molecular evolution. *Proc. Natl Acad. Sci.*

- USA **85**, 2653–2657 (1988).
186. Wolfe, K. H., Sharp, P. M. & Li, W. H. Mutation rates differ among regions of the mammalian genome. *Nature* **337**, 283–285 (1989).
187. Bains, W. Local sequence dependence of rate of base replacement in mammals. *Mutat. Res.* **267**, 43–54 (1992).
188. Mathews, C. K. & Ji, J. DNA precursor asymmetries, replication fidelity, and variable genome evolution. *Bioessays* **14**, 295–301 (1992).
189. Holmquist, G. P. & Filipksi, J. Organization of mutations along the genome: a prime determinant of genome evolution. *Trends Ecol. Evol.* **9**, 65–68 (1994).
190. Eyre-Walker, A. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* **152**, 675–683 (1999).
191. The International SNP Map Working Group. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
192. Bohossian, H. B., Skaletsky, H. & Page, D. C. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature* **406**, 622–625 (2000).
193. Skowronski, J., Fanning, T. G. & Singer, M. F. Unit-length LINE-1 transcripts in human teratocarcinoma cells. *Mol. Cell. Biol.* **8**, 1385–1397 (1988).
194. Boissinot, S., Chevret, P. & Furano, A. V. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol. Biol. Evol.* **17**, 915–928 (2000).
195. Moran, J. V. Human L1 retrotransposition: insights and peculiarities learned from a cultured cell retrotransposition assay. *Genetica* **107**, 39–51 (1999).
196. Kazazian, H. H. Jr *et al.* Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164–166 (1988).
197. Sheen, F.-m. *et al.* Reading between the LINEs: Human genomic variation introduced by LINE-1 retrotransposition. *Genome Res.* **10**, 1496–1508 (2000).
198. Dombroski, B. A., Mathias, S. L., Nanthakumar, E., Scott, A. F. & Kazazian, H. H. Jr Isolation of an active human transposable element. *Science* **254**, 1805–1808 (1991).
199. Holmes, S. E., Dombroski, B. A., Krebs, C. M., Boehm, C. D. & Kazazian, H. H. Jr A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nature Genet.* **7**, 143–148 (1994).
200. Sassaman, D. M. *et al.* Many human L1 elements are capable of retrotransposition. *Nature Genet.* **16**, 37–43 (1997).
201. Dombroski, B. A., Scott, A. F. & Kazazian, H. H. Jr Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proc. Natl Acad. Sci. USA* **90**, 6513–6517 (1993).
202. Kimberland, M. L. *et al.* Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. *Hum. Mol. Genet.* **8**, 1557–1560 (1999).
203. Moran, J. V. *et al.* High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917–927 (1996).
204. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. Jr Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534 (1999).
205. Pickeral, O. K., Makalowski, W., Boguski, M. S. & Boeke, J. D. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* **10**, 411–415 (2000).
206. Miki, Y. *et al.* Disruption of the APC gene by a retrotransposon insertion of L1 sequence in a colon cancer. *Cancer Res.* **52**, 643–645 (1992).
207. Branciforte, D. & Martin, S. L. Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. *Mol. Cell. Biol.* **14**, 2584–2592 (1994).
208. Trelogan, S. A. & Martin, S. L. Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. *Proc. Natl Acad. Sci. USA* **92**, 1520–1524 (1995).
209. Jurka, J. & Kapitonov, V. V. Sectorial mutagenesis by transposable elements. *Genetica* **107**, 239–248 (1999).
210. Fraser, M. J., Ciszczon, T., Elick, T. & Bauser, C. Precise excision of TTAA-specific lepidopteran transposons piggyBac (IP2) and tagalong (TFP3) from the baculovirus genome in cell lines from two species of Lepidoptera. *Insect Mol. Biol.* **5**, 141–151 (1996).
211. Brosius, J. Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* **107**, 209–238 (1999).
212. Kruglyak, S., Durrett, R. T., Schug, M. D. & Aquadro, C. F. Equilibrium distribution of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl Acad. Sci. USA* **95**, 10774–10778 (1998).
213. Toth, G., Gaspari, Z. & Jurka, J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* **10**, 967–981 (2000).
214. Ellegren, H. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature Genet.* **24**, 400–402 (2000).
215. Ji, Y., Eichler, E. E., Schwartz, S. & Nicholls, R. D. Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res.* **10**, 597–610 (2000).
216. Eichler, E. E. Masquerading repeats: paralogous pitfalls of the human genome. *Genome Res.* **8**, 758–762 (1998).
217. Mazzarella, R. & D. Schlessinger, D. Pathological consequences of sequence duplications in the human genome. *Genome Res.* **8**, 1007–1021 (1998).
218. Eichler, E. E. *et al.* Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity. *Hum. Mol. Genet.* **6**, 991–1002 (1997).
219. Horvath, J. E., Schwartz, S. & Eichler, E. E. The mosaic structure of human pericentromeric DNA: a strategy for characterizing complex regions of the human genome. *Genome Res.* **10**, 839–852 (2000).
220. Brand-Arpon, V. *et al.* A genomic region encompassing a cluster of olfactory receptor genes and a myosin light chain kinase (MYLK) gene is duplicated on human chromosome regions 3q13-q21 and 3p13. *Genomics* **56**, 98–110 (1999).
221. Arnold, N., Wienberg, J., Ermet, K. & Zachau, H. G. Comparative mapping of DNA probes derived from the V kappa immunoglobulin gene regions on human and great ape chromosomes by fluorescence in situ hybridization. *Genomics* **26**, 147–150 (1995).
222. Eichler, E. E. *et al.* Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum. Mol. Genet.* **5**, 899–912 (1996).
223. Potier, M. *et al.* Two sequence-ready contigs spanning the two copies of a 200-kb duplication on human 21q: partial sequence and polymorphisms. *Genomics* **51**, 417–426 (1998).
224. Regnier, V. *et al.* Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggest a process of pericentromeric interchromosomal transposition. *Hum. Mol. Genet.* **6**, 9–16 (1997).
225. Ritchie, R. J., Mattei, M. G. & Lalande, M. A large polymorphic repeat in the pericentromeric region of human chromosome 15q contains three partial gene duplications. *Hum. Mol. Genet.* **7**, 1253–1260 (1998).
226. Trask, B. J. *et al.* Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet.* **7**, 13–26 (1998).
227. Trask, B. J. *et al.* Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Hum. Mol. Genet.* **7**, 2007–2020 (1998).
228. van Deutekom, J. C. *et al.* Identification of the first gene (FRG1) from the FSHD region on human chromosome 4q35. *Hum. Mol. Genet.* **5**, 581–590 (1996).
229. Zachau, H. G. The immunoglobulin kappa locus—or—what has been learned from looking closely at one-tenth of a percent of the human genome. *Gene* **135**, 167–173 (1993).
230. Zimonjic, D. B., Kelley, M. J., Rubin, J. S., Aaronson, S. A. & Popescu, N. C. Fluorescence in situ hybridization analysis of keratinocyte growth factor gene amplification and dispersion in evolution of great apes and humans. *Proc. Natl Acad. Sci. USA* **94**, 11461–11465 (1997).
231. van Geel, M. *et al.* The FSHD region on human chromosome 4q35 contains potential coding regions among pseudogenes and a high density of repeat elements. *Genomics* **61**, 55–65 (1999).
232. Horvath, J. E. *et al.* Molecular structure and evolution of an alpha satellite/non-alpha satellite junction at 16p11. *Hum. Mol. Genet.* **9**, 113–123 (2000).
233. Guy, J. *et al.* Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q. *Hum. Mol. Genet.* **9**, 2029–2042 (2000).
234. Reiter, L. T., Murakami, T., Koeuth, T., Gibbs, R. A. & Lupski, J. R. The human COX10 gene is disrupted during homologous recombination between the 24 kb proximal and distal CMT1A-REPs. *Hum. Mol. Genet.* **6**, 1595–1603 (1997).
235. Amos-Landgraf, J. M. *et al.* Chromosome breakage in the Prader-Willi and Angelman syndromes involves recombination between large, transcribed repeats at proximal and distal breakpoints. *Am. J. Hum. Genet.* **65**, 370–386 (1999).
236. Christian, S. L., Fantes, J. A., Mewborn, S. K., Huang, B. & Ledbetter, D. H. Large genomic duplicons map to sites of instability in the Prader-Willi/Angelman syndrome chromosome region (15q11-q13). *Hum. Mol. Genet.* **8**, 1025–1037 (1999).
237. Edelmann, L., Pandita, R. K. & Morrow, B. E. Low-copy repeats mediate the common 3-Mb deletion in patients with velo-cardio-facial syndrome. *Am. J. Hum. Genet.* **64**, 1076–1086 (1999).
238. Shaikh, T. H. *et al.* Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum. Mol. Genet.* **9**, 489–501 (2000).
239. Francke, U. Williams-Beuren syndrome: genes and mechanisms. *Hum. Mol. Genet.* **8**, 1947–1954 (1999).
240. Peoples, R. *et al.* A physical map, including a BAC/PAC clone contig, of the Williams-Beuren syndrome-deletion region at 7q11.23. *Am. J. Hum. Genet.* **66**, 47–68 (2000).
241. Eichler, E. E., Archidiacono, N. & Rocchi, M. CAGG repeats and the pericentromeric duplication of the hominoid genome. *Genome Res.* **9**, 1048–1058 (1999).
242. O'Keefe, C. & Eichler, E. in *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families* (eds Sankoff, D. & Nadeau, J.) 29–46 (Kluwer Academic, Dordrecht, 2000).
243. Lander, E. S. The new genomics: Global views of biology. *Science* **274**, 536–539 (1996).
244. Eddy, S. R. Noncoding RNA genes. *Curr. Opin. Genet. Dev.* **9**, 695–699 (1999).
245. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution. *Science* **289**, 905–920 (2000).
246. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920–930 (2000).
247. Weinstein, L. B. & Steitz, J. A. Guided tours: from precursor snoRNA to functional snoRNP. *Curr. Opin. Cell Biol.* **11**, 378–384 (1999).
248. Bachelier, J.-P. & Cavaillie, J. in *Modification and Editing of RNA* (ed. Benne, H. G. a. R.) 255–272 (ASM, Washington DC, 1998).
249. Burge, C. & Sharp, P. A. Classification of introns: U2-type or U12-type. *Cell* **91**, 875–879 (1997).
250. Brown, C. J. *et al.* The Human Xist gene—analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**, 527–542 (1992).
251. Kickhoefer, V. A., Vasu, S. K. & Rome, L. H. Vaults are the answer, what is the question? *Trends Cell Biol.* **6**, 174–178 (1996).
252. Hatlen, L. & Attardi, G. Proportion of the HeLa cell genome complementary to the transfer RNA and 5S RNA. *J. Mol. Biol.* **56**, 535–553 (1971).
253. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. & Steinberg, S. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **26**, 148–153 (1998).
254. Long, E. O. & Dawid, I. B. Repeated genes in eukaryotes. *Annu. Rev. Biochem.* **49**, 727–764 (1980).
255. Crick, F. H. Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.* **19**, 548–555 (1966).
256. Guthrie, C. & Abelson, J. in *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression* (eds Strathern, J. & Broach, J.) 487–528 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1982).
257. Soll, D. & Rajbandary, U. (eds) *tRNA: Structure, Biosynthesis, and Function* (ASM, Washington DC, 1995).
258. Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**, 13–34 (1985).
259. Bulmer, M. Coevolution of codon usage and transfer-RNA abundance. *Nature* **325**, 728–730 (1987).
260. Duret, L. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* **16**, 287–289 (2000).
261. Sharp, P. M. & Matassi, G. Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* **4**, 851–860 (1994).
262. Buckland, R. A. A primate transfer-RNA gene cluster and the evolution of human chromosome 1. *Cytogenet. Cell Genet.* **61**, 1–4 (1992).
263. Gonos, E. S. & Goddard, J. P. Human tRNA-Glu genes: their copy number and organization. *FEBS Lett.* **276**, 138–142 (1990).
264. Sylvester, J. E. *et al.* The human ribosomal RNA genes: structure and organization of the complete repeating unit. *Hum. Genet.* **73**, 193–198 (1986).

265. Sorensen, P. D. & Frederiksen, S. Characterization of human 5S ribosomal RNA genes. *Nucleic Acids Res.* **19**, 4147–4151 (1991).
266. Timofeeva, M. *et al.* [Organization of a 5S ribosomal RNA gene cluster in the human genome]. *Mol. Biol. (Mosk.)* **27**, 861–868 (1993).
267. Little, R. D. & Braaten, D. C. Genomic organization of human 5S rDNA and sequence of one tandem repeat. *Genomics* **4**, 376–383 (1989).
268. Madsen, B. E. H. The numerous modified nucleotides in eukaryotic ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.* **39**, 241–303 (1990).
269. Tycowski, K. T., You, Z. H., Graham, P. J. & Steitz, J. A. Modification of U6 spliceosomal RNA is guided by other small RNAs. *Mol. Cell* **2**, 629–638 (1998).
270. Pavelitz, T., Liao, D. Q. & Weiner, A. M. Concerted evolution of the tandem array encoding primate U2 snRNA (the RNU2 locus) is accompanied by dramatic remodeling of the junctions with flanking chromosomal sequences. *EMBO J.* **18**, 3783–3792 (1999).
271. Lindgren, V., Ares, A., Weiner, A. M. & Francke, U. Human genes for U2 small nuclear RNA map to a major adenovirus 12 modification site on chromosome 17. *Nature* **314**, 115–116 (1985).
272. Van Arsdell, S. W. & Weiner, A. M. Human genes for U2 small nuclear RNA are tandemly repeated. *Mol. Cell. Biol.* **4**, 492–499 (1984).
273. Gao, L. I., Frey, M. R. & Matera, A. G. Human genes encoding U3 snRNA associate with coiled bodies in interphase cells and are clustered on chromosome 17p11.2 in a complex inverted repeat structure. *Nucleic Acids Res.* **25**, 4740–4747 (1997).
274. Hawkins, J. D. A survey on intron and exon lengths. *Nucleic Acids Res.* **16**, 9893–9908 (1988).
275. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
276. Labeit, S. & Kolmerer, B. Titins: giant proteins in charge of muscle ultrastructure and elasticity. *Science* **270**, 293–296 (1995).
277. Sterner, D. A., Carlo, T. & Berget, S. M. Architectural limits on split genes. *Proc. Natl Acad. Sci. USA* **93**, 15081–15085 (1996).
278. Sun, Q., Mayeda, A., Hampson, R. K., Krainer, A. R. & Rottman, F. M. General splicing factor SF2/ASF promotes alternative splicing by binding to an exonic splicing enhancer. *Genes Dev.* **7**, 2598–2608 (1993).
279. Tanaka, K., Watakabe, A. & Shimura, Y. Polypurine sequences within a downstream exon function as a splicing enhancer. *Mol. Cell. Biol.* **14**, 1347–1354 (1994).
280. Carlo, T., Sterner, D. A. & Berget, S. M. An intron splicing enhancer containing a G-rich repeat facilitates inclusion of a vertebrate micro-exon. *RNA* **2**, 342–353 (1996).
281. Burset, M., Seledtsov, I. A. & Solov'yev, V. V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28**, 4364–4375 (2000).
282. Burge, C. B., Padgett, R. A. & Sharp, P. A. Evolutionary fates and origins of U12-type introns. *Mol. Cell* **2**, 773–785 (1998).
283. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. Frequent alternative splicing of human genes. *Genome Res.* **9**, 1288–1293 (1999).
284. Hanke, J. *et al.* Alternative splicing of human genes: more the rule than the exception? *Trends Genet.* **15**, 389–390 (1999).
285. Brett, D. *et al.* EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* **474**, 83–86 (2000).
286. Dunham, I. The gene guessing game. *Yeast* **17**, 218–224 (2000).
287. Lewin, B. *Gene Expression* (Wiley, New York, 1980).
288. Lewin, B. *Genes IV* 466–481 (Oxford Univ. Press, Oxford, 1990).
289. Smaglik, P. Researchers take a gamble on the human genome. *Nature* **405**, 264 (2000).
290. Fields, C., Adams, M. D., White, O. & Venter, J. C. How many genes in the human genome? *Nature Genet.* **7**, 345–346 (1994).
291. Liang, F. *et al.* Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.* **25**, 239–240 (2000).
292. Roest Crolihus, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25**, 235–238 (2000).
293. The C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**, 2012–2018 (1998).
294. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
295. Green, P. *et al.* Ancient conserved regions in new gene sequences and the protein databases. *Science* **259**, 1711–1716 (1993).
296. Fraser, A. G. *et al.* Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* **408**, 325–330 (2000).
297. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13**, 477–478 (1997).
298. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**, 967–974 (1998).
299. Bailey, L. C. Jr, Searls, D. B. & Overton, G. C. Analysis of EST-driven gene annotation in human genomic sequence. *Genome Res.* **8**, 362–376 (1998).
300. Birney, E., Thompson, J. D. & Gibson, T. J. PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.* **24**, 2730–2739 (1996).
301. Gelfand, M. S., Mironov, A. A. & Pevzner, P. A. Gene recognition via spliced sequence alignment. *Proc. Natl Acad. Sci. USA* **93**, 9061–9066 (1996).
302. Kulp, D., Haussler, D., Reese, M. G. & Eeckman, F. H. A generalized hidden Markov model for the recognition of human genes in DNA. *ISMB* **4**, 134–142 (1996).
303. Reese, M. G., Kulp, D., Tammana, H. & Haussler, D. Genie—gene finding in *Drosophila melanogaster*. *Genome Res.* **10**, 529–538 (2000).
304. Solov'yev, V. & Salamov, A. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *ISMB* **5**, 294–302 (1997).
305. Guigo, R., Agarwal, P., Abril, J. F., Burset, M. & Fickett, J. W. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* **10**, 1631–1642 (2000).
306. Hubbard, T. & Birney, E. Open annotation offers a democratic solution to genome sequencing. *Nature* **403**, 825 (2000).
307. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **28**, 263–266 (2000).
308. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–548 (2000).
309. The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–690 (2001).
310. Basrai, M. A., Hieter, P. & Boeke, J. D. Small open reading frames: beautiful needles in the haystack. *Genome Res.* **7**, 768–771 (1997).
311. Janin, J. & Chothia, C. Domains in proteins: definitions, location, and structural principles. *Methods Enzymol.* **115**, 420–430 (1985).
312. Ponting, C. P., Schultz, J., Copley, R. R., Andrade, M. A. & Bork, P. Evolution of domain families. *Adv. Protein Chem.* **54**, 185–244 (2000).
313. Doolittle, R. F. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64**, 287–314 (1995).
314. Bateman, A. & Birney, E. Searching databases to find protein domain organization. *Adv. Protein Chem.* **54**, 137–157 (2000).
315. Futreal, P. A. *et al.* Cancer and genomics. *Nature* **409**, 850–852 (2001).
316. Nestler, E. J. & Landsman, D. Learning about addiction from the human draft genome. *Nature* **409**, 834–835 (2001).
317. Tupler, R., Perini, G. & Green, M. R. Expressing the human genome. *Nature* **409**, 832–835 (2001).
318. Fahrner, A. M., Bazan, J. F., Papathanasiou, P., Nelms, K. A. & Goodnow, C. C. A genomic view of immunology. *Nature* **409**, 836–838 (2001).
319. Li, W.-H., Gu, Z., Wang, H. & Nekrutenko, A. Evolutionary analyses of the human genome. *Nature* **409**, 847–849 (2001).
320. Bock, J. B., Matern, H. T., Peden, A. A. & Scheller, R. H. A genomic perspective on membrane compartment organization. *Nature* **409**, 839–841 (2001).
321. Pollard, T. D. Genomics, the cytoskeleton and motility. *Nature* **409**, 842–843 (2001).
322. Murray, A. W. & Marks, D. Can sequencing shed light on cell cycling? *Nature* **409**, 844–846 (2001).
323. Clayton, J. D., Kyriacou, C. P. & Reppert, S. M. Keeping time with the human genome. *Nature* **409**, 829–831 (2001).
324. Chervitz, S. A. *et al.* Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* **282**, 2022–2028 (1998).
325. Aravind, L. & Subramanian, G. Origin of multicellular eukaryotes—insights from proteome comparisons. *Curr. Opin. Genet. Dev.* **9**, 688–694 (1999).
326. Attwood, T. K. *et al.* PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.* **28**, 225–227 (2000).
327. Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27**, 215–219 (1999).
328. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
329. Wolf, Y. I., Kondrashov, F. A. & Koonin, E. V. No footprints of primordial introns in a eukaryotic genome. *Trends Genet.* **16**, 333–334 (2000).
330. Brunner, H. G., Nelen, M., Breakefield, X. O., Ropers, H. H. & van Oost, B. B. A. Abnormal behavior associated with a point mutation in the structural gene for monoamine oxidase A. *Science* **262**, 578–580 (1993).
331. Cases, O. *et al.* Aggressive behavior and altered amounts of brain serotonin and norepinephrine in mice lacking MAOA. *Science* **268**, 1763–1766 (1995).
332. Brunner, H. G. *et al.* X-linked borderline mental retardation with prominent behavioral disturbance: phenotype, genetic localization, and evidence for disturbed monoamine metabolism. *Am. J. Hum. Genet.* **52**, 1032–1039 (1993).
333. Deckert, J. *et al.* Excess of high activity monoamine oxidase A gene promoter alleles in female patients with panic disorder. *Hum. Mol. Genet.* **8**, 621–624 (1999).
334. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
335. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
336. Ponting, C. P., Aravind, L., Schultz, J., Bork, P. & Koonin, E. V. Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J. Mol. Biol.* **289**, 729–745 (1999).
337. Zhang, J., Dyer, K. D. & Rosenberg, H. F. Evolution of the rodent eosinophil-associated RNase gene family by rapid gene sorting and positive selection. *Proc. Natl Acad. Sci. USA* **97**, 4701–4706 (2000).
338. Shashoua, V. E. Ependymin, a brain extracellular glycoprotein, and CNS plasticity. *Ann. NY Acad. Sci.* **627**, 94–114 (1991).
339. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28**, 231–234 (2000).
340. Koonin, E. V., Aravind, L. & Kondrashov, A. S. The impact of comparative genomics on our understanding of evolution. *Cell* **101**, 573–576 (2000).
341. Bateman, A., Eddy, S. R. & Chothia, C. Members of the immunoglobulin superfamily in bacteria. *Protein Sci.* **5**, 1939–1941 (1996).
342. Sutherland, D., Samakolis, C. & Krasnow, M. A. Branchless encodes a *Drosophila* FGF homolog that controls tracheal cell migration and the pattern of branching. *Cell* **87**, 1091–1101 (1996).
343. Warburton, D. *et al.* The molecular basis of lung morphogenesis. *Mech. Dev.* **92**, 55–81 (2000).
344. Fuchs, T., Glusman, G., Horn-Saban, S., Lancet, D. & Pilpel, Y. The human olfactory subgenome: from sequence to structure to evolution. *Hum. Genet.* **108**, 1–13 (2001).
345. Glusman, G. *et al.* The olfactory receptor gene family: data mining, classification and nomenclature. *Mamm. Genome* **11**, 1016–1023 (2000).
346. Rouquier, S. *et al.* Distribution of olfactory receptor genes in the human genome. *Nature Genet.* **18**, 243–250 (1998).
347. Sharon, D. *et al.* Primate evolution of an olfactory receptor cluster: Diversification by gene conversion and recent emergence of a pseudogene. *Genomics* **61**, 24–36 (1999).
348. Gilad, Y. *et al.* Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. *Nature Genet.* **26**, 221–224 (2000).
349. Gearhart, J. & Kirschner, M. *Cells, Embryos, and Evolution* (Blackwell Science, Malden, Massachusetts, 1997).
350. Barbazuk, W. B. *et al.* The syntenic relationship of the zebrafish and human genomes. *Genome Res.* **10**, 1351–1358 (2000).
351. McLysaght, A., Enright, A. J., Skrabanek, L. & Wolfe, K. H. Estimation of syntenic conservation and genome compaction between pufferfish (*Fugu*) and human. *Yeast* **17**, 22–36 (2000).
352. Trachtulec, Z. *et al.* Linkage of TATA-binding protein and proteasome subunit C5 genes in mice and humans reveals syntenic conserved between mammals and invertebrates. *Genomics* **44**, 1–7 (1997).

353. Nadeau, J. H. Maps of linkage and syntenic homologies between mouse and man. *Trends Genet.* **5**, 82–86 (1989).
354. Nadeau, J. H. & Taylor, B. A. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl Acad. Sci. USA* **81**, 814–818 (1984).
355. Copeland, N. G. *et al.* A genetic linkage map of the mouse: current applications and future prospects. *Science* **262**, 57–66 (1993).
356. DeBry, R. W. & Seldin, M. F. Human/mouse homology relationships. *Genomics* **33**, 337–351 (1996).
357. Nadeau, J. H. & Sankoff, D. The lengths of undiscovered conserved segments in comparative maps. *Mamm. Genome* **9**, 491–495 (1998).
358. Thomas, J. W. *et al.* Comparative genome mapping in the sequence-based era: early experience with human chromosome 7. *Genome Res.* **10**, 624–633 (2000).
359. Pletcher, M. T. *et al.* Chromosome evolution: The junction of mammalian chromosomes in the formation of mouse chromosome 10. *Genome Res.* **10**, 1463–1467 (2000).
360. Novacek, M. J. Mammalian phylogeny: shaking the tree. *Nature* **356**, 121–125 (1992).
361. O'Brien, S. J. *et al.* Genome maps 10. Comparative genomics. Mammalian radiations. Wall chart. *Science* **286**, 463–478 (1999).
362. Romer, A. S. *Vertebrate Paleontology* (Univ. Chicago Press, Chicago and New York, 1966).
363. Paterson, A. H. *et al.* Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. *Nature Genet.* **14**, 380–382 (1996).
364. Jenczewski, E., Prosper, J. M. & Ronfort, J. Differentiation between natural and cultivated populations of *Medicago sativa* (Leguminosae) from Spain: analysis with random amplified polymorphic DNA (RAPD) markers and comparison to allozymes. *Mol. Ecol.* **8**, 1317–1330 (1999).
365. Ohno, S. *Evolution by Gene Duplication* (George Allen and Unwin, London, 1970).
366. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).
367. Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. Extensive duplication and reshuffling in the arabidopsis genome. *Plant Cell* **12**, 1093–1102 (2000).
368. Paterson, A. H. *et al.* Comparative genomics of plant chromosomes. *Plant Cell* **12**, 1523–1540 (2000).
369. Vision, T., Brown, D. & Tanksley, S. The origins of genome duplications in *Arabidopsis*. *Science* **290**, 2114–2117 (2000).
370. Sidow, A. & Bowman, B. H. Molecular phylogeny. *Curr. Opin. Genet. Dev.* **1**, 451–456 (1991).
371. Sidow, A. & Thomas, W. K. A molecular evolutionary framework for eukaryotic model organisms. *Curr. Biol.* **4**, 596–603 (1994).
372. Sidow, A. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6**, 715–722 (1996).
373. Spring, J. Vertebrate evolution by interspecific hybridisation—are we polyploid? *FEBS Lett.* **400**, 2–8 (1997).
374. Skrabanek, L. & Wolfe, K. H. Eukaryote genome duplication—where's the evidence? *Curr. Opin. Genet. Dev.* **8**, 694–700 (1998).
375. Hughes, A. L. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.* **48**, 565–576 (1999).
376. Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
377. Horikawa, Y. *et al.* Genetic variability in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genet.* **26**, 163–175 (2000).
378. Hastbacka, J. *et al.* The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell* **78**, 1073–1087 (1994).
379. Tischkoff, S. A. *et al.* Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387 (1996).
380. Kidd, J. R. *et al.* Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus PAH, in a global representation of populations. *Am. J. Hum. Genet.* **63**, 1882–1899 (2000).
381. Mateu, E. *et al.* Worldwide genetic analysis of the CFTR region. *Am. J. Hum. Genet.* **68**, 103–117 (2001).
382. Abecasis, G. R. *et al.* Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68**, 191–197 (2001).
383. Taillon-Miller, P. *et al.* Juxtaposed regions of extensive and minimal linkage disequilibrium in Xq25 and Xq28. *Nature Genet.* **25**, 324–328 (2000).
384. Martin, E. R. *et al.* SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *Am. J. Hum. Genet.* **67**, 383–394 (2000).
385. Collins, A., Lonjou, C. & Morton, N. E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96**, 15173–15177 (1999).
386. Dunning, A. M. *et al.* The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.* **67**, 1544–1554 (2000).
387. Rieder, M. J., Taylor, S. L., Clark, A. G. & Nickerson, D. A. Sequence variation in the human angiotensin converting enzyme. *Nature Genet.* **22**, 59–62 (1999).
388. Collins, F. S. Positional cloning moves from perditional to traditional. *Nature Genet.* **9**, 347–350 (1995).
389. Nagamine, K. *et al.* Positional cloning of the APECED gene. *Nature Genet.* **17**, 393–398 (1997).
390. Reuber, B. E. *et al.* Mutations in PEX1 are the most common cause of peroxisome biogenesis disorders. *Nature Genet.* **17**, 445–448 (1997).
391. Portsteffen, H. *et al.* Human PEX1 is mutated in complementation group 1 of the peroxisome biogenesis disorders. *Nature Genet.* **17**, 449–452 (1997).
392. Everett, L. A. *et al.* Pendred syndrome is caused by mutations in a putative sulphate transporter gene (PDS). *Nature Genet.* **17**, 411–422 (1997).
393. Coffey, A. J. *et al.* Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. *Nature Genet.* **20**, 129–135 (1998).
394. Van Laer, L. *et al.* Nonsyndromic hearing impairment is associated with a mutation in DFNA5. *Nature Genet.* **20**, 194–197 (1998).
395. Sakuntabhai, A. *et al.* Mutations in ATP2A2, encoding a Ca²⁺ pump, cause Darier disease. *Nature Genet.* **21**, 271–277 (1999).
396. Gedeon, A. K. *et al.* Identification of the gene (SEDL) causing X-linked spondyloepiphyseal dysplasia tarda. *Nature Genet.* **22**, 400–404 (1999).
397. Hurvitz, J. R. *et al.* Mutations in the CCN gene family member WISP3 cause progressive pseudorheumatoid dysplasia. *Nature Genet.* **23**, 94–98 (1999).
398. Laberge-le Couteux, S. *et al.* Truncating mutations in CCM1, encoding KRIT1, cause hereditary cavernous angiomas. *Nature Genet.* **23**, 189–193 (1999).
399. Sahoo, T. *et al.* Mutations in the gene encoding KRIT1, a Krev-1/rap1a binding protein, cause cerebral cavernous malformations (CCM1). *Hum. Mol. Genet.* **8**, 2325–2333 (1999).
400. McGuirt, W. T. *et al.* Mutations in COL11A2 cause non-syndromic hearing loss (DFNA13). *Nature Genet.* **23**, 413–419 (1999).
401. Moreira, E. S. *et al.* Limb-girdle muscular dystrophy type 2G is caused by mutations in the gene encoding the sarcomeric protein telethonin. *Nature Genet.* **24**, 163–166 (2000).
402. Ruiz-Perez, W. L. *et al.* Mutations in a new gene in Ellis-van Creveld syndrome and Weyers acrodermal dysostosis. *Nature Genet.* **24**, 283–286 (2000).
403. Kaplan, J. M. *et al.* Mutations in ACTN4, encoding alpha-actinin-4, cause familial focal segmental glomerulosclerosis. *Nature Genet.* **24**, 251–256 (2000).
404. Escayg, A. *et al.* Mutations of SCN1A, encoding a neuronal sodium channel, in two families with GEFS+2. *Nature Genet.* **24**, 343–345 (2000).
405. Sacksteder, K. A. *et al.* Identification of the alpha-aminoacidic semialdehyde synthase gene, which is defective in familial hyperlysinemia. *Am. J. Hum. Genet.* **66**, 1736–1743 (2000).
406. Kalaydjieva, L. *et al.* N-myc downstream-regulated gene 1 is mutated in hereditary motor and sensory neuropathy-Lom. *Am. J. Hum. Genet.* **67**, 47–58 (2000).
407. Sundin, O. H. *et al.* Genetic basis of total colourblindness among the Pingelapese islanders. *Nature Genet.* **25**, 289–293 (2000).
408. Kohl, S. *et al.* Mutations in the CNGB3 gene encoding the beta-subunit of the cone photoreceptor cGMP-gated channel are responsible for achromatopsia (ACHM3) linked to chromosome 8q21. *Hum. Mol. Genet.* **9**, 2107–2116 (2000).
409. Avela, K. *et al.* Gene encoding a new RING-B-box-coiled-coil protein is mutated in mulibrey nanism. *Nature Genet.* **25**, 298–301 (2000).
410. Verpy, E. *et al.* A defect in harmonin, a PDZ domain-containing protein expressed in the inner ear sensory hair cells, underlies usher syndrome type 1C. *Nature Genet.* **26**, 51–55 (2000).
411. Bitner-Glindzic, M. *et al.* A recessive contiguous gene deletion causing infantile hyperinsulinism, enteropathy and deafness identifies the usher type 1C gene. *Nature Genet.* **26**, 56–60 (2000).
412. The May-Hegglin/Fetchnier Syndrome Consortium. Mutations in MYH9 result in the May-Hegglin anomaly, and Fechtner and Sebastian syndromes. *Nature Genet.* **26**, 103–105 (2000).
413. Kelley, M. J., Jawien, W., Ortel, T. L. & Korczak, J. F. Mutation of MYH9, encoding non-muscle myosin heavy chain A, in May-Hegglin anomaly. *Nature Genet.* **26**, 106–108 (2000).
414. Kirschner, L. S. *et al.* Mutations of the gene encoding the protein kinase A type I-α regulatory subunit in patients with the Carney complex. *Nature Genet.* **26**, 89–92 (2000).
415. Lalwani, A. K. *et al.* Human nonsyndromic hereditary deafness DFNA17 is due to a mutation in non-muscle myosin MYH9. *Am. J. Hum. Genet.* **67**, 1121–1128 (2000).
416. Matsuura, T. *et al.* Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nature Genet.* **26**, 191–194 (2000).
417. Delettre, C. *et al.* Nuclear gene OPA1, encoding a mitochondrial dynamin-related protein, is mutated in dominant optic atrophy. *Nature Genet.* **26**, 207–210 (2000).
418. Pusch, C. M. *et al.* The complete form of X-linked congenital stationary night blindness is caused by mutations in a gene encoding a leucine-rich repeat protein. *Nature Genet.* **26**, 324–327 (2000).
419. The ADHR Consortium. Autosomal dominant hypophosphataemic rickets is associated with mutations in FGF23. *Nature Genet.* **26**, 345–348 (2000).
420. Bomont, P. *et al.* The gene encoding gigaxonin, a new member of the cytoskeletal BTB/kelch repeat family, is mutated in giant axonal neuropathy. *Nature Genet.* **26**, 370–374 (2000).
421. Tullio-Pelet, A. *et al.* Mutant WD-repeat protein in triple-A syndrome. *Nature Genet.* **26**, 332–335 (2000).
422. Nicole, S. *et al.* Perlecan, the major proteoglycan of basement membranes, is altered in patients with Schwartz-Jampel syndrome (chondrodystrophic myotonia). *Nature Genet.* **26**, 480–483 (2000).
423. Rogaev, E. I. *et al.* Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature* **376**, 775–778 (1995).
424. Sherrington, R. C. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **375**, 754–760 (1995).
425. Olivieri, N. F. & Weatherall, D. J. The therapeutic reactivation of fetal haemoglobin. *Hum. Mol. Genet.* **7**, 1655–1658 (1998).
426. Drews, J. Research & development. Basic science and pharmaceutical innovation. *Nature Biotechnol.* **17**, 406 (1999).
427. Drews, J. Drug discovery: a historical perspective. *Science* **287**, 1960–1964 (2000).
428. Davies, P. A. *et al.* The 5-HT3B subunit is a major determinant of serotonin-receptor function. *Nature* **397**, 359–363 (1999).
429. Heise, C. E. *et al.* Characterization of the human cysteinyl leukotriene 2 receptor. *J. Biol. Chem.* **275**, 30531–30536 (2000).
430. Fan, W. *et al.* BACE maps to chromosome 11 and a BACE homolog, BACE2, reside in the obligate Down Syndrome region of chromosome 21. *Science* **286**, 1255a (1999).
431. Saunders, A. J., Kim, T. -W. & Tanzi, R. E. BACE maps to chromosome 11 and a BACE homolog, BACE2, reside in the obligate Down Syndrome region of chromosome 21. *Science* **286**, 1255a (1999).
432. Firestein, S. The good taste of genomics. *Nature* **404**, 552–553 (2000).
433. Matsunami, H., Montmayeur, J. P. & Buck, L. B. A family of candidate taste receptors in human and mouse. *Nature* **404**, 601–604 (2000).
434. Adler, E. *et al.* A novel family of mammalian taste receptors. *Cell* **100**, 693–702 (2000).
435. Chandrashekar, J. *et al.* T2Rs function as bitter taste receptors. *Cell* **100**, 703–711 (2000).
436. Hardison, R. C. Conserved non-coding sequences are reliable guides to regulatory elements. *Trends Genet.* **16**, 369–372 (2000).
437. Onyango, P. *et al.* Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain. *Genome Res.* **10**, 1697–1710 (2000).
438. Bouck, J. B., Metzker, M. L. & Gibbs, R. A. Shotgun sample sequence comparisons between mouse and human genomes. *Nature Genet.* **25**, 31–33 (2000).
439. Marshall, E. Public-private project to deliver mouse genome in 6 months. *Science* **290**, 242–243 (2000).
440. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.* **26**, 225–228 (2000).
441. Tagle, D. A. *et al.* Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203**, 439–455 (1988).
442. McGuire, A. M., Hughes, J. D. & Church, G. M. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* **10**, 744–757 (2000).

443. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.* **16**, 939–945 (1998).
444. Cheng, Y. & Church, G. M. Biclustering of expression data. *ISMB* **8**, 93–103 (2000).
445. Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.* **26**, 183–186 (2000).
446. Feil, R. & Khosla, S. Genomic imprinting in mammals: an interplay between chromatin and DNA methylation? *Trends Genet.* **15**, 431–434 (1999).
447. Robertson, K. D. & Wolffe, A. P. DNA methylation in health and disease. *Nature Rev. Genet.* **1**, 11–19 (2000).
448. Beck, S., Olek, A. & Walter, J. From genomics to epigenomics: a loftier view of life. *Nature Biotechnol.* **17**, 1144–1144 (1999).
449. Hagmann, M. Mapping a subtext in our genetic book. *Science* **288**, 945–946 (2000).
450. Eliot, T. S. in *T. S. Eliot. Collected Poems 1909–1962* (Harcourt Brace, New York, 1963).
451. Soderland, C., Longden, I. & Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13**, 523–535 (1997).
452. Mott, R. & Tribe, R. Approximate statistics of gapped alignments. *J. Comp. Biol.* **6**, 91–112 (1999).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

Beyond the authors, many people contributed to the success of this work. E. Jordan provided helpful advice throughout the sequencing effort. We thank D. Leja and J. Shehadeh for their expert assistance on the artwork in this paper, especially the foldout figure; K. Jegalian for editorial assistance; J. Schloss, E. Green and M. Seldin for comments on an earlier version of the manuscript; P. Green and F. Ouellette for critiques of the submitted version; C. Caulcott, A. Iglesias, S. Renfrey, B. Skene and J. Stewart of the Wellcome Trust, P. Whittington and T. Dougans of NHGRI and M. Meunier of Genoscope for staff support for meetings of the international consortium; and the University of Pennsylvania for facilities for a meeting of the genome analysis group.

We thank Compaq Computer Corporation's High Performance Technical Computing Group for providing a Compaq Biocluster (a 27 node configuration of AlphaServer ES40s, containing 108 CPUs, serving as compute nodes and a file server with one terabyte of secondary storage) to assist in the annotation and analysis. Compaq provided the systems and implementation services to set up and manage the cluster for continuous use by

members of the sequencing consortium. Platform Computing Ltd. provided its LSF scheduling and loadsharing software without license fee.

In addition to the data produced by the members of the International Human Genome Sequencing Consortium, the draft genome sequence includes published and unpublished human genomic sequence data from many other groups, all of whom gave permission to include their unpublished data. Four of the groups that contributed particularly significant amounts of data were: M. Adams *et al.* of the Institute for Genomic Research; E. Chen *et al.* of the Center for Genetic Medicine and Applied Biosystems; S.-F. Tsai of National Yang-Ming University, Institute of Genetics, Taipei, Taiwan, Republic of China; and Y. Nakamura, K. Koyama *et al.* of the Institute of Medical Science, University of Tokyo, Human Genome Center, Laboratory of Molecular Medicine, Minato-ku, Tokyo, Japan. Many other groups provided smaller numbers of database entries. We thank them all; a full list of the contributors of unpublished sequence is available as Supplementary Information.

This work was supported in part by the National Human Genome Research Institute of the US NIH; The Wellcome Trust; the US Department of Energy, Office of Biological and Environmental Research, Human Genome Program; the UK MRC; the Human Genome Sequencing Project from the Science and Technology Agency (STA) Japan; the Ministry of Education, Science, Sport and Culture, Japan; the French Ministry of Research; the Federal German Ministry of Education, Research and Technology (BMBF) through Projektträger DLR, in the framework of the German Human Genome Project; BEO, Projektträger Biologie, Energie, Umwelt des BMBF und BMWt; the Max-Planck-Society; DFG—Deutsche Forschungsgemeinschaft; TMWFK, Thüringer Ministerium für Wissenschaft, Forschung und Kunst; EC BIOMED2—European Commission, Directorate Science, Research and Development; Chinese Academy of Sciences (CAS), Ministry of Science and Technology (MOST), National Natural Science Foundation of China (NSFC); US National Science Foundation EPSCoR and The SNP Consortium Ltd. Additional support for members of the Genome Analysis group came, in part, from an ARCS Foundation Scholarship to T.S.F., a Burroughs Wellcome Foundation grant to C.B.B. and P.A.S., a DFG grant to P.B., DOE grants to D.H., E.E.E. and T.S.F., an EU grant to P.B., a Marie-Curie Fellowship to L.C., an NIH-NHGRI grant to S.R.E., an NIH grant to E.E.E., an NIH SBIR to D.K., an NSF grant to D.H., a Swiss National Science Foundation grant to L.C., the David and Lucille Packard Foundation, the Howard Hughes Medical Institute, the University of California at Santa Cruz and the W. M. Keck Foundation.

Correspondence and requests for materials should be addressed to E. S. Lander (e-mail: lander@genome.wi.mit.edu), R. H. Waterston (e-mail: bwatrst@watson.wustl.edu), J. Sulston (e-mail: jes@sanger.ac.uk) or F. S. Collins (e-mail: fc23a@nih.gov).

Affiliations for authors: 1, Whitehead Institute for Biomedical Research, Center for Genome Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA; 2, The Sanger Centre, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1RQ, United Kingdom; 3, Washington University Genome Sequencing Center, Box 8501, 4444 Forest Park Avenue, St. Louis, Missouri 63108, USA; 4, US DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA; 5, Baylor College of Medicine Human Genome Sequencing Center, Department of Molecular and Human Genetics, One Baylor Plaza, Houston, Texas 77030, USA; 6, Department of Cellular and Structural Biology, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, Texas 78229-3900, USA; 7, Department of Molecular Genetics, Albert Einstein College of Medicine, 1635 Poplar Street, Bronx, New York 10461, USA; 8, Baylor College of Medicine Human Genome Sequencing Center and the Department of Microbiology & Molecular Genetics, University of Texas Medical School, PO Box 20708, Houston, Texas 77225, USA; 9, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku Yokohama-city, Kanagawa 230-0045, Japan; 10, Genoscope and CNRS UMR-8030, 2 Rue Gaston Cremieux, CP 5706, 91057 Evry Cedex, France; 11, GTC Sequencing Center, Genome Therapeutics Corporation, 100 Beaver Street, Waltham, Massachusetts 02453-8443, USA; 12, Department of Genome Analysis, Institute of Molecular Biotechnology, Beutenbergstrasse 11, D-07745 Jena, Germany; 13, Beijing Genomics Institute/Human Genome Center, Institute of Genetics, Chinese Academy of Sciences, Beijing 100101, China; 14, Southern China National Human Genome Research Center, Shanghai 201203, China; 15, Northern China National Human Genome Research Center, Beijing 100176, China; 16, Multimegabase Sequencing Center, The Institute for Systems Biology, 4225 Roosevelt Way, NE Suite 200, Seattle, Washington 98105, USA; 17, Stanford Genome Technology Center, 855 California Avenue, Palo Alto, California 94304, USA; 18, Stanford Human Genome Center and Department of Genetics, Stanford University School of Medicine, Stanford, California 94305-5120, USA; 19, University of Washington Genome Center, 225 Fluke Hall on Mason Road, Seattle, Washington 98195, USA; 20, Department of Molecular Biology, Keio University School of Medicine, 35 Shinanomachi, Shinjuku-ku, Tokyo 160-8582, Japan; 21, University of Texas Southwestern Medical Center at Dallas, 6000 Harry Hines Blvd., Dallas, Texas 75235-8591, USA; 22, University of

Oklahoma's Advanced Center for Genome Technology, Dept. of Chemistry and Biochemistry, University of Oklahoma, 620 Parrington Oval, Rm 311, Norman, Oklahoma 73019, USA; 23, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany; 24, Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA; 25, GBF - German Research Centre for Biotechnology, Mascheroder Weg 1, D-38124 Braunschweig, Germany; 26, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bldg. 38A, 8600 Rockville Pike, Bethesda, Maryland 20894, USA; 27, Department of Genetics, Case Western Reserve School of Medicine and University Hospitals of Cleveland, BRB 720, 10900 Euclid Ave., Cleveland, Ohio 44106, USA; 28, EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom; 29, Max Delbrück Center for Molecular Medicine, Robert-Rossle-Strasse 10, 13125 Berlin-Buch, Germany; 30, EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany; 31, Dept. of Biology, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, Massachusetts 02139-4307, USA; 32, Howard Hughes Medical Institute, Dept. of Genetics, Washington University School of Medicine, Saint Louis, Missouri 63110, USA; 33, Dept. of Computer Science, University of California at Santa Cruz, Santa Cruz, California 95064, USA; 34, Affymetrix, Inc., 2612 8th St, Berkeley, California 94710, USA; 35, Genome Exploration Research Group, Genomic Sciences Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan; 36, Howard Hughes Medical Institute, Department of Computer Science, University of California at Santa Cruz, California 95064, USA; 37, University of Dublin, Trinity College, Department of Genetics, Smurfit Institute, Dublin 2, Ireland; 38, Cambridge Research Laboratory, Compaq Computer Corporation and MIT Genome Center, 1 Cambridge Center, Cambridge, Massachusetts 02142, USA; 39, Dept. of Mathematics, University of California at Santa Cruz, Santa Cruz, California 95064, USA; 40, Dept. of Biology, University of California at Santa Cruz, Santa Cruz, California 95064, USA; 41, Crown Human Genetics Center and Department of Molecular Genetics, The Weizmann Institute of Science, Rehovot 71600, Israel; 42, Dept. of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA; 43, The University of Michigan Medical School, Departments of Human Genetics and Internal Medicine, Ann Arbor, Michigan

48109, USA; 44, MRC Functional Genetics Unit, Department of Human Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK; 45, Institute for Systems Biology, 4225 Roosevelt Way NE, Seattle, WA 98105, USA; 46, National Human Genome Research Institute, US National Institutes of Health, 31 Center Drive, Bethesda, Maryland 20892, USA; 47, Office of Science, US Department of Energy, 19901 Germantown Road, Germantown, Maryland 20874, USA; 48, The Wellcome Trust, 183 Euston Road, London, NW1 2BE, UK.

† Present addresses: Genome Sequencing Project, Egea Biosciences, Inc., 4178 Sorrento Valley Blvd., Suite F, San Diego, CA 92121, USA (G.A.E.); INRA, Station d'Amélioration des Plantes, 63039 Clermont-Ferrand Cedex 2, France (L.C.).

DNA sequence databases

GenBank, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bldg. 38A, 8600 Rockville Pike, Bethesda, Maryland 20894, USA

EMBL, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

DNA Data Bank of Japan, Center for Information Biology, National Institute of Genetics, 1111 Yata, Mishima-shi, Shizuoka-ken 411-8540, Japan

Experimental annotation of the human genome using microarray technology

D. D. Shoemaker*, E. E. Schadt*, C. D. Armour, Y. D. He, P. Garrett-Engle, P. D. McDonagh, P. M. Loerch, A. Leonardson, P. Y. Lum, G. Cavet, L. F. Wu, S. J. Altschuler, S. Edwards, J. King, J. S. Tsang, G. Schimmack, J. M. Schelter, J. Koch, M. Ziman, M. J. Marton, B. Li, P. Cundiff, T. Ward, J. Castle, M. Krolewski, M. R. Meyer, M. Mao, J. Burchard, M. J. Kidd, H. Dai, J. W. Phillips, P. S. Linsley, R. Stoughton, S. Scherer & M. S. Boguski

Rosetta Inpharmatics, Inc., 12040 115th Avenue N.E., Kirkland, Washington 98034, USA

* These authors contributed equally to this work

The most important product of the sequencing of a genome is a complete, accurate catalogue of genes and their products, primarily messenger RNA transcripts and their cognate proteins. Such a catalogue cannot be constructed by computational annotation alone; it requires experimental validation on a genome scale. Using 'exon' and 'tiling' arrays fabricated by ink-jet oligonucleotide synthesis, we devised an experimental approach to validate and refine computational gene predictions and define full-length transcripts on the basis of co-regulated expression of their exons. These methods can provide more accurate gene numbers and allow the detection of mRNA splice variants and identification of the tissue- and disease-specific conditions under which genes are expressed. We apply our technique to chromosome 22q under 69 experimental condition pairs, and to the entire human genome under two experimental conditions. We discuss implications for more comprehensive, consistent and reliable genome annotation, more efficient, full-length complementary DNA cloning strategies and application to complex diseases.

The initial interpretation of a genome sequence rests upon conclusions derived solely from bioinformatics approaches—*ab initio* gene predictions, homology studies, motif analysis and other non-experimental methods^{1–3}. The limitations and fallibility of this process have been discussed^{4,5} and one group has concluded⁶ that, despite more than 17 years of research effort⁷, precise annotation of every gene in the human genome by computational methods alone is still a distant goal. Bioinformatics analyses of fragmentary experimental data have led to widely varying estimates of the number of human genes^{8–10}. Comparative genomics approaches, particularly between human and mouse^{11–13}, will help to identify candidate genes and refine their structures, but cannot alone show that a gene is active. Consequently, projects to clone and catalogue 'full-length' cDNA clones from human¹⁴ and mouse¹⁵ have been undertaken. Although these projects may capture the complete coding sequences of many genes in time, cDNA cloning fixes a gene product at a particular time and under particular conditions, and thus cannot efficiently reveal the multifunctional nature of a metazoan transcriptome.

Recent work indicates that the human genome may contain fewer genes than anticipated^{8,9}, and that frequent alternative splicing might account for much physiological complexity^{16–19}. This situation makes it essential to pursue a course that efficiently yields empirical validation of the structures of genes and simultaneously provides an accurate and complete catalogue of their expressed products (mRNA and cognate protein sequences).

We describe a high-throughput, microarray-based experimental method to validate predicted exons, group the exons into genes by co-regulated expression and define full-length mRNA transcripts. The method involves the design and fabrication of 'exon arrays' consisting of long (50–60 bases) oligonucleotide probes derived from predicted exons, followed by hybridization with fluorescently labelled cDNAs derived from specific cell lines or normal or diseased tissues. Absolute intensities (measuring cellular abundances) or intensity ratios (measuring differential expression regulation) from hybridized cDNAs are used to identify those probes that represent authentic exons under the conditions tested. In addition, the expression data can define gene boundaries, because adjacent exons that are co-regulated across many conditions are likely to be from the same transcript. For a higher-resolution view of gene

structure, we use 'tiling arrays' in which overlapping oligonucleotides are designed to blanket an entire genomic region of interest. This approach can potentially reveal exons not identified by current gene prediction algorithms and provide information about alternative splicing.

We applied the exon array approach to a detailed analysis of human chromosome 22 under 69 pairs of experimental conditions. Tiling arrays were used to refine the structure of new genes discovered by exon analysis. Finally, a preliminary analysis of the entire human genome using exon arrays under two experimental conditions demonstrated the power of being able experimentally to validate hundreds of thousands of exon predictions, anticipating the prospect of analysing the entire human genome to a depth similar to that achieved on chromosome 22.

Analysis of chromosome 22q using exon arrays

Chromosome 22 was the first human chromosome to be completely sequenced and subjected to exhaustive computational annotation². It has thus served as a benchmark for new computational and experimental methods of analysis^{20,21}. We designed a single ink-jet array to monitor the 8,183 exons annotated² on chromosome 22q under diverse experimental conditions. Specifically, mRNAs from human cell lines and normal and diseased tissues (Fig. 1) were fluorescently labelled with two colours and hybridized in pairs to 69 individual chromosome 22 exon arrays (see Methods). Figure 2a shows a graphical display of error-weighted log expression ratios²² for all 8,183 exons across 69 condition pairs. We developed a gene identification algorithm that uses intensity and ratio information to identify exons in a local neighbourhood that are strongly correlated across condition pairs, and then to extend such regions by incorporating other local exons with similar expression behaviour. The resulting 572 groups of co-regulated exons are referred to as expression-verified genes (EVGs). Figure 2b–e shows expanded views of specific regions of chromosome 22. Expression data can be used to confirm the exons and structure of a known gene (Fig. 2b), to identify potential false positive exon predictions (Fig. 2c), to merge UniGene clusters into a single gene (Fig. 2d) and to verify *ab initio* gene predictions experimentally (Fig. 2e).

For a chromosome-wide performance summary, we compared our experimentally derived EVGs to the list of 545 genes annotated

by Dunham *et al.*² (Table 1). These annotated genes were divided into four categories (known, related, predicted and *ab initio*) on the basis of the level of experimental support for the predictions. We identified 210 (85%) of the 247 known genes by analysing the expression data from the 69 condition pairs with our gene detection algorithm. The remaining 15% of known genes did not exhibit sufficient differential expression regulation among the conditions tested to enable ratio-based algorithms to verify them. We detected 66% of the related genes and 53% of the predicted genes using our expression regulation criteria. The most interesting result comes from the 325 *ab initio* genes that represented pure Genscan predictions. Dunham *et al.*² speculated that only 100 of these predicted transcripts would represent portions of 'real' genes, but we found experimental support for 185 (57%) of the genes in this category.

A few of the EVGs that we detected contained more than one gene. This occurred when adjacent genes were co-regulated across the 69 experimental conditions tested. In most cases, this situation can be addressed by testing additional conditions or by using additional bioinformatics techniques (for example, open reading frame (ORF) analysis, identification of internal polyadenylation sites, and supporting expressed sequence tag (EST) and protein sequence data). In a few cases, a single gene was represented by more than one EVG, indicating possible alternative splicing. Other algorithms are being developed to address this issue.

Applications of tiling arrays

Exon-based gene validation arrays can be limited by the fact that gene prediction programs perform best on 'internal' exons and not very well on initial and terminal exons, or exons that correspond to the 5' and 3' untranslated regions (UTRs) of mRNAs⁶. Oligonucleotide tiling arrays of overlapping probes (Fig. 3) can effectively address this challenge because they are constructed without any *a priori* knowledge of the possible exon content of a genomic sequence. We designed tiling arrays covering both strands of various genomic regions on chromosome 22 defined by EVGs where the underlying gene structure was thought to be incomplete.

Figure 3 shows how the tiling approach was used to refine the structure of the novel testis transcript described in Fig. 2e. We fabricated an ink-jet array that contained 60-mer probes spaced in

10-base-pair (bp) intervals across both strands of the 113-kilobase (kb) bacterial artificial chromosome (BAC) clone containing the EVG of interest. The array was hybridized with fluorescently labelled testis mRNA and the resulting probe intensities were analysed to determine the approximate locations of the exons within this region. For each exon, the hybridization data effectively reduced the search for the intron–exon boundaries to regions of around 20–30 bp. The exact splice junctions can generally be identified within these narrow windows by using common rules (for example, GT-AG consensus sequence and ORF analysis). For the gene shown in Fig. 3, only four of the six exons were correctly predicted by Genscan. Our results extend the 3' UTR by 450 bp and one of the internal coding exons by 102 bases (34 amino acids). These results were confirmed by polymerase chain reaction with reverse transcriptase (RT-PCR) and sequencing (data not shown). The mRNA (GenBank accession no. AF324466) derived from this validated and corrected gene is 1,312 nucleotides long, including a 649-base 3' UTR with a polyadenylation signal at base 1,293. It encodes a 217-residue protein and a BLASTP search revealed only one significant match (*E*-value $\sim 10^{-15}$) to a predicted gene product, CG5280 from the *Drosophila* genome project²³.

Human genome scan using exon arrays

To show that the approach described above can scale to survey the entire human genome, we used the 15 June 2000 version of the Ensembl human genome annotation data set (<http://www.ensembl.org/>)²⁴ to make 50 arrays containing 1,090,408 oligonucleotide probes representing 442,785 exons predicted by Genscan²⁵. Fluorescently labelled cDNAs from a human lymphoma cell line and a colorectal carcinoma cell line were hybridized to the arrays. Analysis of fluorescence intensities from this single pair of experimental conditions provided experimental evidence for 58% of the 78,486 Ensembl confirmed exons. We detected 34% of the 364,299 predicted exons that did not meet the Ensembl 'confirmed' criteria. The false positive rate for this analysis was estimated to be around 5%, from an analysis of a set of negative control probes included on the arrays. A summary of the exons validated by this genome survey is given (see Methods) in Fig. 4 and a full listing is available as Supplementary Information or from the Rosetta website at www.rii.com.

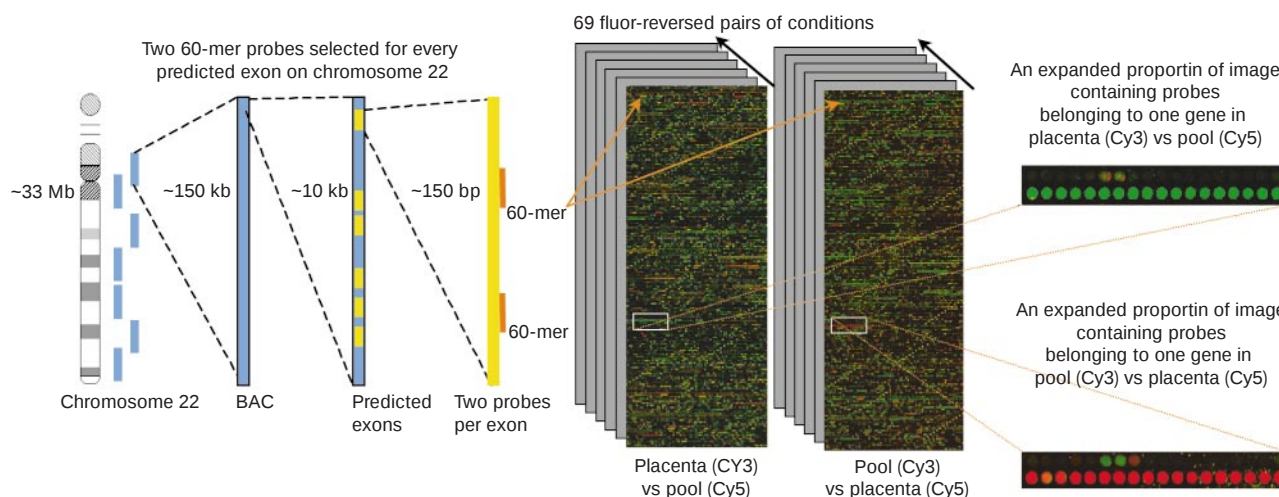


Figure 1 Design and fabrication of exon arrays for the predicted exons on human chromosome 22. Two 60-mers were selected from each of 8,183 predicted exons on human chromosome 22q and printed on a single 1 x 3 inch array (~25,000 60-mers). This array was hybridized with 69 pairs of RNA samples using a two-colour hybridization technique. Each experiment was performed in duplicate with a fluor reversal to minimize

possible bias caused by the molecular structure of the Cy3 and Cy5 dyes (138 arrays in total). Red and green spots, as shown in the expanded panels on the right, are probes representing experimentally verified genes (groups of differentially expressed exons that are located next to each other in the genome).



Figure 2 Using expression data from multiple conditions to validate exons and define gene boundaries on chromosome 22. **a**, Pseudocolour image showing error-weighted log₁₀ expression ratios (red/green) for each of the ~8,000 exons (x-axis) across the 69 fluor-reversed experiments (y-axis). A brief description of each experiment is listed on the right side of the image; the numbers (1–69) are reference points for the Table in the Supplementary Information. The 15,511 probes representing the 8,183 predicted exons are arranged linearly across the 33 Mb of chromosome 22. **b**, Expanded region showing a known gene (SERPIND1, NM_000185). The experiments on the y-axis have been clustered to emphasize how co-regulation across diverse experiments can be used to

group exons into genes. The vertical white lines indicate the boundaries predicted by our gene finding algorithm; numbers on y-axis indicate experimental conditions. **c**, Expanded region showing a set of co-regulated exons from another known gene (G22P1, NM_001469), illustrating the detection of potential false positives (arrow) made by the Genscan prediction program. **d**, Expanded region representing an EVG that collapses two Unigene EST clusters (HS.269963 and HS.14587) into a single transcript. **e**, Expanded region showing an EVG containing six exons that are part of a novel testis-expressed transcript (arrows, two experiments involving testis RNA samples).

Discussion

Post-genome biology and medicine will increasingly rely on complete and accurate catalogues of human genes, mRNAs and proteins. This 'parts list' is currently a patchwork of mostly hypothetical entities with varying degrees of supporting evidence. Computational techniques for sequence annotation provide invaluable clues to gene structure and function but experimental data will be required to provide a full and satisfying picture. Our microarray-based technology represents a comprehensive and consistent

approach to the simultaneous validation of gene predictions and study of the transcriptome under any number of biologically or medically interesting conditions. Our approach is applicable on a genome scale and also on the scale of defining the structure of a single, novel cDNA.

The exon-based approach is well suited to high-throughput screening of diverse cell types, growth conditions and disease states. Differential expression is an important tool for assembling exons into genes. We could detect differential expression for only 15% of the confirmed exons across the human genome with a single condition pair. Clearly, larger data sets will be essential for defining the structures of genes, detecting rarely expressed genes and addressing more complex issues such as alternative splicing. In addition, information from the exon analysis can be used to select genomic regions and samples for comprehensive tiling arrays.

Ambitious efforts to clone and sequence 'full-length' cDNAs for the human¹⁴ and mouse¹⁵ genomes have begun with the purpose not only of helping to validate computational gene predictions but also of providing physical reagents for functional and structural geno-

Table 1 Gene validation summary of human chromosome 22q

	Annotation from ref. 2	Expression-verified genes (EVGs)	Validation fraction
Known genes*	247	210	85%
Related genes*	150	99	66%
Predicted genes*	148	78	53%
<i>Ab initio</i> genes*	325	185	57%

EVG sequences were searched against current versions of dbEST and nr (www.ncbi.nlm.nih.gov) and significant matches were defined as those having an *E*-value $< 10^{-20}$.

* Category definitions according to Dunham *et al.*².

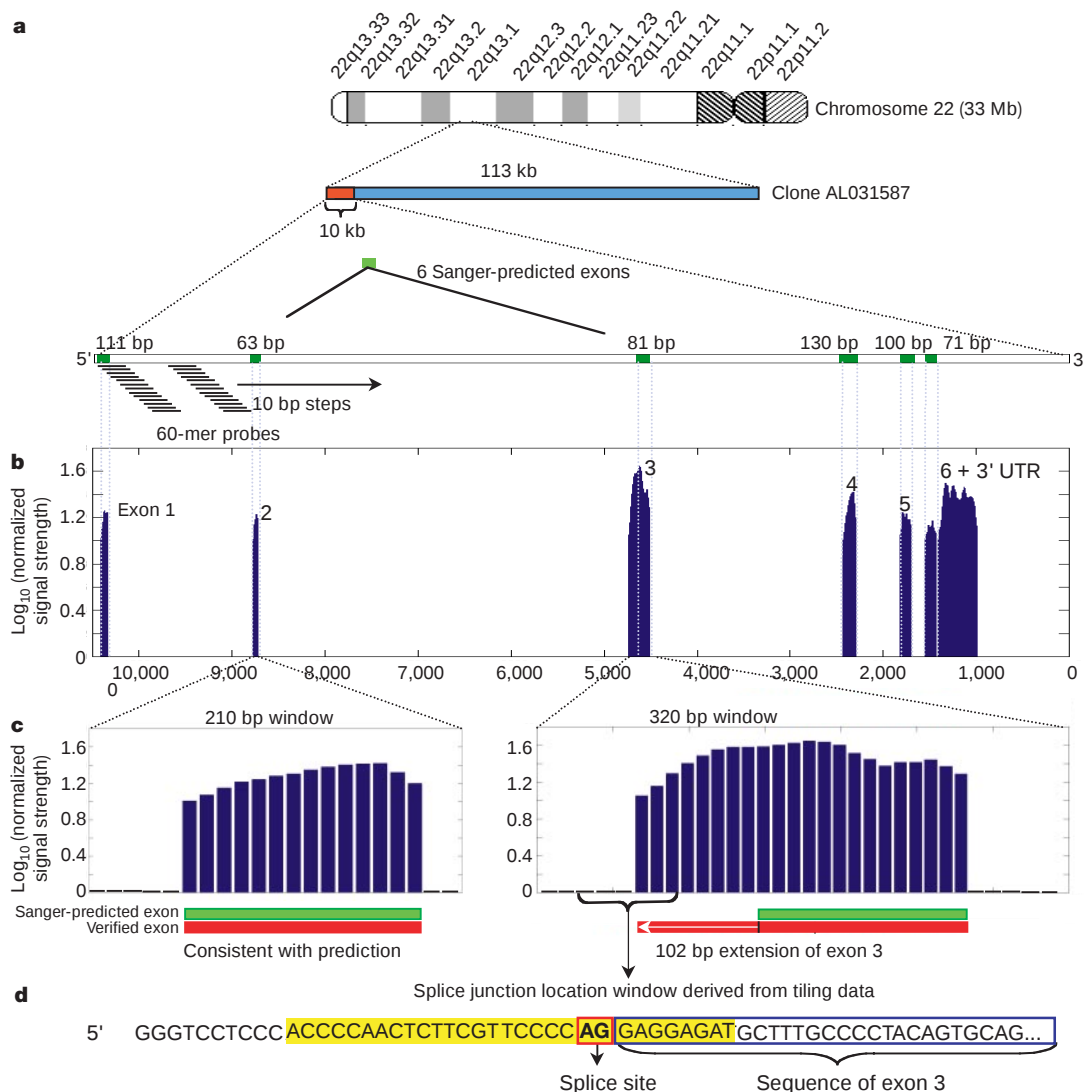


Figure 3 Characterization of a novel testis transcript using tiling arrays. **a**, An EVG discovered in the analysis of chromosome 22 (Fig. 2e) was localized to a 10-kb region at one end of the insert of BAC clone AL031587. Both strands of this 113-kb genomic interval were tiled with 60-mer probes at 10-bp steps. The tiling array was hybridized with RNA from human testis. **b**, Hybridization signals corresponding to tiling probes from this region were filtered and plotted as $\log_{10}$ values of the normalized signal strengths. Of the

six Genscan predicted exons in this region, two (exons 3 and 6) were at variance with the hybridization data. **c**, Detailed views of tiling data showing one correctly predicted exon and one incorrectly predicted exon. **d**, Typically, tiling data narrow the search window for an intron/exon boundary to 20–30-bp. The exact splice junction is then identified using consensus sequences (GT-AG rule) and ORF information. The exact splice junction can also be determined by sequencing RT–PCR products.

mics. The comprehensive set of EVGs generated by our approach will accelerate these efforts by allowing a more directed cloning strategy. We also expect that hybridization data defining EVGs will be useful in 'training' the next generation of gene prediction algorithms, in much the same manner that sequence similarity data enhances *ab initio* predictions in the current state-of-the-art programs. In this way, the maximum value can be realized from the intersection of computational and high-throughput experimental biology.

Our experimental method of annotating the human genome could be rapidly reiterated for updated sequence information from the Human Genome Project, and could easily be extended to the genomes of other organisms. Generating exon and tiling arrays requires only the availability of genomic sequence and exon predictions, from which probes can be rapidly and efficiently synthesized onto an array. The flexibility and short time scale for designing

and fabricating exon and tiling arrays using the ink-jet platform could substantially accelerate gene discovery.

Finally, our approach could be useful in the identification and analysis of genes underlying complex diseases. Genetic linkage studies of polygenic traits typically yield a dozen loci, each up to 20–30 megabases long. It is feasible to construct tiling arrays across all loci and probe them with mRNA samples from relevant normal and diseased tissues to ascertain both gene content and activity. Such analyses may provide not only more direct routes to the culpable genes, but also have the potential to uncover regulatory mutations by observed alterations in gene activity. □

Methods

Sources of predicted exons

To analyse chromosome 22q, we designed a single ink-jet oligonucleotide microarray to represent 8,183 sequences that had been identified or confirmed as having coding potential (Sanger Centre). We used two sources of information: 6,650 Genscan-predicted exon sequences, and 3,381 validated exon sequences identified by aligning the first complete version of the human chromosome 22 sequence with sequences from experimentally validated transcripts². Of this set of 10,031 exons, 1,847 had coordinates identical to those of other exons and were removed from the sequence pool. The remaining 8,183 exon sequences were subjected to an oligonucleotide design process to identify the two best candidate probes for a given exon sequence (see below). For the whole-genome exon scan, we designed ink-jet oligonucleotide microarrays to 442,785 predicted exons selected from the publicly available assembled sequence in the Ensembl database as of 15 June 2000. Specifically, we selected 554,202 non-redundant sequences from an initial set of 628,635 Genscan predicted exons²⁴. We removed 111,417 more sequences from the list after they were flagged by the RepeatMasker algorithm (<http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>).

Probe selection for the exon-scanning arrays

For each of the predicted exons, we selected the top two 60-mers using an algorithm that takes into account binding energies, base composition, sequence complexity, cross-hybridization binding energies and secondary structure. For exon sequences of 60 nucleotides or less, we designed a single probe consisting of the entire exon sequence. For the 8,183 predicted exons on chromosome 22, 15,511 60-mers were selected and printed on a single array. For the whole-genome exon arrays, we selected 1,090,408 60mers to represent the 442,785 GenScan predicted exons from the Ensembl database. For 78,486 of the exons annotated as 'confirmed', the reverse-complement probes were also selected and placed next to the regular probes on the array as negative controls.

Probe selection for tiling arrays

In the tiling experiment described in Fig. 3, 60-mer probes were placed in 10-bp intervals across a 113.8-kb region of chromosome 22 containing the novel testis transcript described in Fig. 2e (BAC clone AL031587). The reverse complements for each of the tiling probes were also included on the array to allow transcripts on either strand to be detected. The genomic sequences used in the tiling experiments were repeat-masked before probe selection but no other exclusionary filters were applied.

Array synthesis

We synthesized the oligonucleotide arrays on 1 × 3-inch glass slides using ink-jet technology²⁶. The phosphoramidite monomers were delivered by a standard ink-jet printer head to defined positions on a glass surface containing exposed hydroxyl groups. The remaining synthesis steps are similar to traditional oligonucleotide synthesis. Using this approach, up to 25,000 different 60-mers can be synthesized on a single slide. Around 1,000 'gridline' probes (5' CCTATGTGACTGGTCGATGCTACTA 3') are placed around the perimeter of each array. Fluorescently labelled synthetic oligonucleotides complementary to the control probes are included in all hybridizations. Arrays based on Rosetta designs were purchased from Agilent Technologies.

Preparation of labelled cDNA

We used the following human cell lines: Jurkat (T lymphocyte, ATCC no. TIB-152), K562 (chronic myelogenous leukaemia, ATCC no. CCL-243), Raji (Burkitt's lymphoma, ATCC no. CCL-86), Colo (colorectal adenocarcinoma, ATCC no. CCL-220), 293 (embryonic kidney, ATCC no. CRL-1573.1) and HepG2 (hepatocellular carcinoma, ATCC no. CRL-11997). Poly-A⁺ RNA (mRNA) was isolated from each of the cytoplasmic RNA samples as described²⁷. The 'pool' RNA sample described in Fig. 2 contains an equal mixture of four human cell lines (Jurkat, K562, Raji and Colo). The 41 mRNA samples from the human tissues described in Fig. 2 were purchased from commercial sources and are described at www.rii.com/Publications. For a single hybridization, we combined 1.5 µg of mRNA with 1.0 µg of random 9-mers and incubated the mixture for 10 min at 70 °C, 5 min at 4 °C and 10 min at 22 °C. To this mixture we added 0.5 mM amino-allyl dUTP (Sigma A-0410), 0.5 mM dNTP, 1 × RT buffer, 5 mM MgCl₂, 10 mM DTT and 200 units of Superscript (GibcoBRL), bringing the final reverse transcription reaction volume to 40 µl. This reverse transcription reaction was incubated for 20 min at 42 °C and the RNA was hydrolysed by adding 20 µl EDTA + NaOH and incubating at 65 °C for 20 min. The reaction was

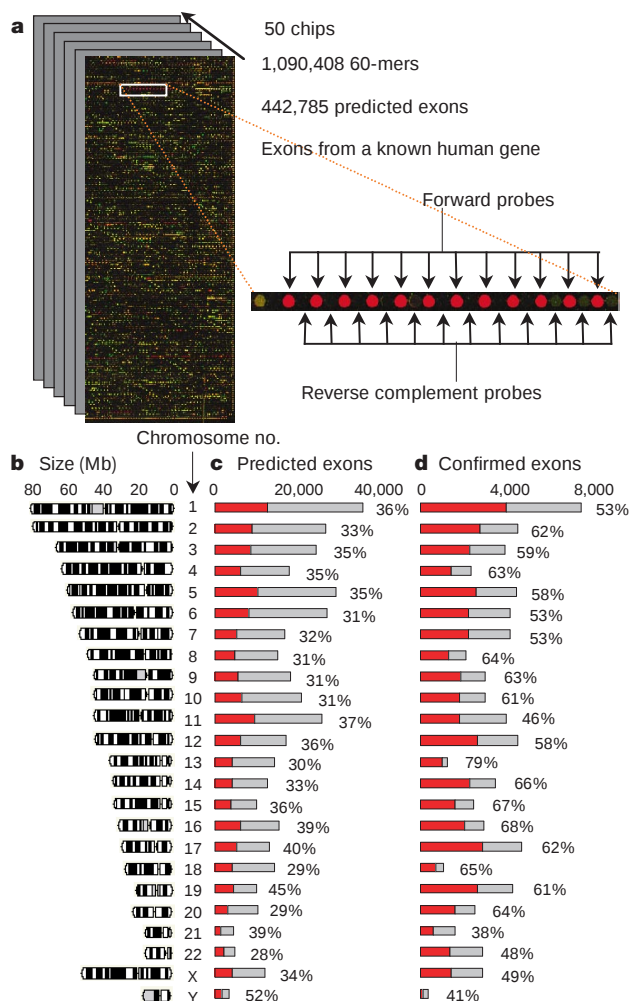


Figure 4 Whole-genome scan for validating predicted exons. **a**, Fifty 1 × 3-inch ink-jet arrays were used to test 442,785 Genscan predicted exons under two conditions. For each predicted exon, the best one or two 60-mer probes were selected, resulting in the set of 1,090,408 probes which were distributed over 50 arrays (~25,000 60-mers per array). The arrays also included 110,000 reverse complement probes and 48,500 control probes. The arrays were hybridized with Cy-3 or Cy-5 labelled mRNA from two human cell lines (Raji and Colo). Enlarged image, probes representing exons from a known gene with alternating forward and reverse complement probes. All experiments were performed in duplicate with a fluor reversal (100 arrays total). **b**, The sizes of the 24 human chromosomes (left). **c**, The number of predicted exons that were experimentally verified (red bars) for each of the chromosomes. Grey bars, number of predicted exons on each chromosome. **d**, A similar analysis for the confirmed exons across the human genome.

neutralized by adding 20 μ l of 1M Tris-HCl pH 7.6. We concentrated the resulting amino-allyl labelled single-stranded cDNA using a Microcon-30 (Millipore), and coupled it to Cy3 or Cy5 dye using a CyDye kit (Amersham Pharmacia Q15108). The per cent dye incorporation and total cDNA yield were determined spectrophotometrically. Pairs of Cy5/Cy3-labelled cDNA samples were combined and hybridized as described²².

Analysis and visual display of expression data

Array images were processed as described²² to obtain background noise, single channel intensity and associated measurement error estimates. Expression changes between two samples were quantified as $\log_{10}$ (expression ratio) where the 'expression ratio' was taken to be the ratio between normalized, background-corrected intensity values for the two channels (red and green) for each spot on the array. An error model for the log ratio was applied²² to quantify the significance of expression changes between two samples. The colour displays in Fig. 2 show $\log_{10}$ (expression ratio) as red when the red channel is upregulated relative to the green channel, green when the red channel is downregulated relative to the green channel, black when $\log_{10}$ (expression ratio) is close to zero, and grey when data from one or both of the channels for a given probe are unreliable.

Identifying EVGs by co-regulation

Exons were grouped into EVGs by a two-step gene identification algorithm. First, each probe was assigned a similarity measure, based on the moving average (using a window size equal to six probes) of pair-wise Pearson correlation coefficients between the log ratios of probe intensities in neighbouring exons. Probes with correlation coefficients above 0.5 in a given window were selected as seeds for EVGs. The 0.5 threshold and window size were determined empirically by training the model on a subset of the known chromosome 22 genes. Second, probes neighbouring a seed region were merged into the region if the pair-wise correlation coefficients between the neighbouring probe and the average in the seed region exceeded 0.5. This process continued, allowing for gaps between probe pairs to account for failed probes and/or false exon predictions (gaps were not allowed to exceed five probes), until no probes flanking the candidate region met the significance threshold of correlation with the exon cluster. The final exon clusters resulting from the gene detection algorithm were identified as an EVG. Not all condition pairs (rows) were considered in forming EVGs. Elements in a given row had to have significant *P* values (≤ 0.01) to be included in the analysis. Once an EVG was formed, the colour display (as in Fig. 2) was updated by reordering the condition pairs according to a hierarchical clustering algorithm, as described²⁸.

Annotation of EVGs

Predicted transcripts for all EVGs identified from the chromosome 22 exon data across the 69 condition pairs were formed by combining the individual exons into a single sequence. Each of these sequences was searched against dbEST and the NR protein databases using gapped BLASTN and BLASTX (www.ncbi.nlm.nih.gov), respectively, to determine the extent to which the EVG sequences were similar to other sequence data. We declared sequences similar if the corresponding *E*-value for the alignment was less than 10^{-9} , using default parameters for gapped BLAST. BLAST results were used to determine the degree of sequence support defining a predicted transcript. These results were also used to determine the degree of existing sequence support for each of the EVGs detected from the chromosome 22 exon arrays.

Quantitative analysis of whole-genome exon data

We used an intensity-based algorithm to verify predicted exons experimentally across the entire human genome. Specifically, we used raw intensity measurements for the forward-strand (FS) probes and the corresponding raw intensity measurements for the reverse-complement (RC) probes in conjunction with the respective standard deviations of those measurements to determine the significance of the FS probe intensities. We controlled for nonspecific cross-hybridization using RC probes, given that the reverse complement of a DNA sequence has equivalent sequence complexity to the forward strand sequence with respect to a variety of measures (such as GC content and GC trend). An exon was called 'present' if the intensity difference between an FS probe and the RC probe had $P < 0.01$ in either the red or green channel, and if the FS probe intensity had a $P < 0.01$ for being above background in the channel where the difference was considered most significant. *P* values were calculated using a *t*-test applied to the difference of the mean pixel intensities and to the difference of the mean FS/background intensities.

These single channel exon detection methods were applied only to those exons in which reverse-complement probes were designed. In the remaining cases, the significance of the single channel intensities was determined using the above-background criterion described above. We applied a correction to the detection percentages given for the predicted exons listed in Fig. 4, based on false positive estimates for above-background calls that were determined using the FS/RC probe intensity difference calls for the confirmed exons. Error

models used in this analysis to assess ratio significance were as described²⁸. Of the 88,374 confirmed exons represented on the genome-wide exon arrays, 78,486 had corresponding RC probes. To assess the rate of false positives expected in the single-channel assessments, we used a similar detection procedure to determine the number of RC probe intensity measurements that were significantly greater than the corresponding FS probe intensity. Our results indicate that the false positive rate of detection using the single channel method was $\sim 5\%$.

Received 28 November 2000; accepted 9 January 2001.

1. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
2. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
3. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
4. Wheelan, S. J. & Boguski, M. S. Late-night thoughts on the sequence annotation problem. *Genome Res.* **8**, 168–169 (1998).
5. Boguski, M. S. Biosequence exegesis. *Science* **286**, 453–455 (1999).
6. Guigo, R., Agarwal, P., Abril, J. F., Burset, M. & Fickett, J. W. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* **10**, 1631–1642 (2000).
7. Claverie, J. M. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6**, 1735–1744 (1997).
8. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25**, 232–234 (2000).
9. Roest Croliis, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetradon nigroviridis* DNA sequence. *Nature Genet.* **25**, 235–238 (2000).
10. Liang, F. *et al.* Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.* **25**, 239–240 (2000).
11. Makalowski, W. & Boguski, M. S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA* **95**, 9407–9412 (1998).
12. Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* **10**, 950–958 (2000).
13. Marshall, E. Public-private project to deliver mouse genome in 6 months. *Science* **290**, 242–243 (2000).
14. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. The mammalian gene collection. *Science* **286**, 455–457 (1999).
15. The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. Functional annotation of a full-length mouse cDNAs collection. *Nature* **409**, 685–690 (2001).
16. Hanke, J. *et al.* Alternative splicing of human genes: more the rule than the exception? *Trends Genet.* **15**, 389–390 (1999).
17. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. Frequent alternative splicing of human genes. *Genome Res.* **9**, 1288–1293 (1999).
18. Black, D. L. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **103**, 367–370 (2000).
19. Brett, D. *et al.* EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* **474**, 83–86 (2000).
20. de Souza, S. J. *et al.* Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc. Natl Acad. Sci. USA* **97**, 12690–12693 (2000).
21. Penn, S. G., Rank, D. R., Hanzel, D. K. & Barker, D. L. Mining the human genome using microarrays of open reading frames. *Nature Genet.* **26**, 315–318 (2000).
22. Roberts, C. J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
23. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
24. Hubbard, T. & Birney, E. Open annotation offers a democratic solution to genome sequencing. *Nature* **403**, 825 (2000).
25. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
26. Blanchard, A. P., Kaiser, R. J. & Hood, L. E. High-density oligonucleotide arrays. *Biosens. Bioelectron.* **6/7**, 687–690 (1996).
27. Marton, M. J. *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* **4**, 1293–1301 (1998).
28. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank S. H. Friend for encouragement and support and J. Rine, M. V. Olson, C. Roberts and T. Hughes for critical readings of the manuscript.

Correspondence and requests for materials should be addressed to M.S.B. (e-mail: msb@rii.com).

A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms

The International SNP Map Working Group*

* A full list of authors appears at the end of this paper.

We describe a map of 1.42 million single nucleotide polymorphisms (SNPs) distributed throughout the human genome, providing an average density on available sequence of one SNP every 1.9 kilobases. These SNPs were primarily discovered by two projects: The SNP Consortium and the analysis of clone overlaps by the International Human Genome Sequencing Consortium. The map integrates all publicly available SNPs with described genes and other genomic features. We estimate that 60,000 SNPs fall within exon (coding and untranslated regions), and 85% of exons are within 5 kb of the nearest SNP. Nucleotide diversity varies greatly across the genome, in a manner broadly consistent with a standard population genetic model of human history. This high-density SNP map provides a public resource for defining haplotype variation across the genome, and should help to identify biomedically important genes for diagnosis and therapy.

Inherited differences in DNA sequence contribute to phenotypic variation, influencing an individual's anthropometric characteristics, risk of disease and response to the environment. A central goal of genetics is to pinpoint the DNA variants that contribute most significantly to population variation in each trait. Genome-wide linkage analysis and positional cloning have identified hundreds of genes for human diseases¹ (<http://ncbi.nlm.nih.gov/OMIM>), but nearly all are rare conditions in which mutation of a single gene is necessary and sufficient to cause disease. For common diseases, genome-wide linkage studies have had limited success, consistent with a more complex genetic architecture. If each locus contributes modestly to disease aetiology, more powerful methods will be required.

One promising approach is systematically to explore the limited set of common gene variants for association with disease^{2–4}. In the human population most variant sites are rare, but the small number of common polymorphisms explain the bulk of heterozygosity³ (see also refs 5–11). Moreover, human genetic diversity appears to be limited not only at the level of individual polymorphisms, but also in the specific combinations of alleles (haplotypes) observed at closely linked sites^{8,11–14}. As these common variants are responsible for most heterozygosity in the population, it will be important to assess their potential impact on phenotypic trait variation.

If limited haplotype diversity is general, it should be practical to define common haplotypes using a dense set of polymorphic markers, and to evaluate each haplotype for association with disease. Such haplotype-based association studies offer a significant advantage: genomic regions can be tested for association without requiring the discovery of the functional variants. The required density of markers will depend on the complexity of the local haplotype structure, and the distance over which these haplotypes extend, neither of which is yet well defined.

Current estimates (refs 13–17) indicate that a very dense marker map (30,000–1,000,000 variants) would be required to perform haplotype-based association studies. Most human sequence variation is attributable to SNPs, with the rest attributable to insertions or deletions of one or more bases, repeat length polymorphisms and rearrangements. SNPs occur (on average) every 1,000–2,000 bases when two human chromosomes are compared^{15,6,9,18–20}, and are thus present at sufficient density for comprehensive haplotype analysis. SNPs are binary, and thus well suited to automated,

high-throughput genotyping. Finally, in contrast to more mutable markers, such as microsatellites²¹, SNPs have a low rate of recurrent mutation, making them stable indicators of human history. We have constructed a SNP map of the human genome with sufficient density to study human haplotype structure, enabling future study of human medical and population genetics.

Identification and characteristics of SNPs

The map contains all SNPs that were publicly available in November 2000. Over 95% were discovered by The SNP Consortium (TSC) and the public Human Genome Project (HGP). TSC contributed 1,023,950 candidate SNPs (<http://snp.cshl.org>) identified by shotgun sequencing of genomic fragments drawn from a complete (45% of data) or reduced (55% of data) representation of the human genome^{18,22}. Individual contributions were: Whitehead Institute, 589,209 SNPs from 2.57 million (M) passing reads; Sanger Centre, 262,279 SNPs from 1.16M passing reads; Washington University, 172,462 SNPs from 1.69M passing reads. TSC SNPs were discovered using a publicly available panel of 24 ethnically diverse individuals²³. Reads were aligned to one another and to the available genome sequence, followed by detection of single base differences using one of two validated algorithms: Polybayes²⁴ and the neighbourhood quality standard (NQS^{18,22}).

An additional 971,077 candidate SNPs were identified as sequence differences in regions of overlap between large-insert clones (bacterial artificial chromosomes (BACs) or P1-derived artificial chromosomes (PACs)) sequenced by the HGP. Two groups (NCBI/Washington University (556,694 SNPs): G.B., P.Y.K. and S.S.; and The Sanger Centre (630,147 SNPs): J.C.M. and D.R.B.) independently analysed these overlaps using the two detection algorithms. This approach contributes dense clusters of SNPs throughout the genome. The remaining 5% of SNPs were discovered in gene-based studies, either by automated detection of single base differences in clusters of overlapping expressed sequence tags^{24–28} or by targeted resequencing efforts (see ftp://ncbi.nlm.nih.gov/snp/human/submit_format/*/*publicat.rep.gz).

It is critical that candidate SNPs have a high likelihood of representing true polymorphisms when examined in population studies. Although many methods and contributors are represented on the map (see above), most SNPs (> 95%) were contributed by two large-scale efforts that uniformly applied automated methods.

Random samples of these SNPs have been evaluated by confirmation in the original DNA samples (where possible) to rule out false positives, and in independent population samples to determine allele frequency. The TSC centres and two outside laboratories (Orchid and Cold Spring Harbor Laboratory) successfully genotyped 1,585 TSC SNPs in the 24 DNA samples used for discovery (<http://snp.cshl.org>); having surveyed all chromosomes in which each SNP could have been identified, any non-polymorphic candidates must represent false positives. In these tests, 1,500 SNPs (95%) were polymorphic, 67 (4%) non-polymorphic (false positives) and 18 (1%) uniformly heterozygous (previously unrecognized repeats). These high validation rates were observed separately for subsets of SNPs discovered by reduced representation shotgun and genomic alignment, and for subsets identified with Polybayes and the NQS. Thus, these algorithms appear to generate few false positive SNPs. The small number (1%) of uniformly 'heterozygous' candidate SNPs show that the methods also exclude nearly all low-copy repeats.

The allele frequencies of a set of SNPs have been evaluated²⁹ in independent populations using pooled resequencing. Samples of TSC ($n = 502$) and overlap SNPs ($n = 774$) were studied in population samples of European, African American and Chinese descent, revealing 82% to be polymorphic in at least one ethnic group at frequencies above the detection threshold of pooled resequencing (~10%). The remaining 18% presumably represent SNPs with a frequency less than 10% in the populations surveyed and false positives. Furthermore, 77% of SNPs had a minor allele frequency of more than 20% in at least one population, and 27% had an allele frequency higher than 20% in all three ethnic groups. TSC and overlap SNPs had similar distributions across the populations, showing that they are comparable in quality and frequency. The high proportion of SNPs with significant population frequency is expected after SNP discovery in two or a few chromosomes, given standard assumptions about human population history^{18,29,30}.

Description of the SNP map

We mapped the sequence flanking each SNP by alignment to the genomic sequence of large-insert clones in Genbank. These alignments were converted into chromosomal coordinates according to

Table 1 SNP distribution by chromosome

Chromosome	Length (bp)	All SNPs		TSC SNPs	
		SNPs	kb per SNP	SNPs	kb per SNP
1	214,066,000	129,931	1.65	75,166	2.85
2	222,889,000	103,664	2.15	76,985	2.90
3	186,938,000	93,140	2.01	63,669	2.94
4	169,035,000	84,426	2.00	65,719	2.57
5	170,954,000	117,882	1.45	63,545	2.69
6	165,022,000	96,317	1.71	53,797	3.07
7	149,414,000	71,752	2.08	42,327	3.53
8	125,148,000	57,834	2.16	42,653	2.93
9	107,440,000	62,013	1.73	43,020	2.50
10	127,894,000	61,298	2.09	42,466	3.01
11	129,193,000	84,663	1.53	47,621	2.71
12	125,198,000	59,245	2.11	38,136	3.28
13	93,711,000	53,093	1.77	35,745	2.62
14	89,344,000	44,112	2.03	29,746	3.00
15	73,467,000	37,814	1.94	26,524	2.77
16	74,037,000	38,735	1.91	23,328	3.17
17	73,367,000	34,621	2.12	19,396	3.78
18	73,078,000	45,135	1.62	27,028	2.70
19	56,044,000	25,676	2.18	11,185	5.01
20	63,317,000	29,478	2.15	17,051	3.71
21	33,824,000	20,916	1.62	9,103	3.72
22	33,786,000	28,410	1.19	11,056	3.06
X	131,245,000	34,842	3.77	20,400	6.43
Y	21,753,000	4,193	5.19	1,784	12.19
RefSeq	15,696,674	14,534	1.08		
Totals	2,710,164,000	1,419,190	1.91	887,450	3.05

Length (bp) is from the public Genome Assembly of 5 September 2000. Density of SNPs on each chromosome is influenced by the amount of available genome sequence included in the Genome Assembly, depth of overlap coverage from TSC reads and clone overlaps, and the underlying heterozygosity (Table 2). Data are presented for the entire dataset (All SNPs) and for those from the SNP consortium (TSC SNPs), as the latter are more evenly spaced than those from clone overlaps.

the publicly available genome assemblies of July and September 2000 (<http://genome.ucsc.edu>). Candidate SNPs were included in the final map only if they mapped to a single location in the genome assembly. Integrated displays of SNPs, genes and other features are available at the ENSEMBL (<http://www.ensembl.org>), NCBI (National Center for Biotechnology Information; <http://www.ncbi.nlm.nih.gov>), UCSC (University of California at Santa Cruz; <http://genome.ucsc.edu>) and TSC (<http://snp.cshl.org>) websites.

The nonredundant SNP total of 1,433,393 is fewer than the sum of individual submissions (2,067,476) because some SNPs (mainly in regions of BAC overlap) were discovered by more than one effort. Of these, 1,419,190 mapped to unique locations in the 2.7 gigabases (Gb) of assembled human genome sequence, providing an average density of one SNP every 1.91 kb. TSC SNPs, which are more evenly distributed than those from clone overlaps, were found on average every 3.05 kb. SNP density (Table 1) is relatively constant across the autosomes. To characterize the distribution of SNPs, we examined 366,192 SNPs that fell within finished sequence. Most of the genome contains SNPs at high density (Fig. 1): 90% of contiguous 20-kb windows contain one or more SNPs, as do 63% of 5-kb windows and 28% of 1-kb windows. Only 4% of genome sequence falls in gaps between SNPs of > 80 kb, and some of these gaps are covered by SNPs that are discovered but not yet mapped owing to gaps in the genome assembly.

To evaluate the density of SNPs in regions within and surrounding genes, we used the September 2000 release of RefSeq³¹. In total, 14,534 SNPs map to within these 7,000 carefully annotated, non-redundant messenger RNAs, equivalent to about two exonic SNPs per gene (coding and untranslated regions). Extrapolating two exonic SNPs per gene to the approximately 30,000 human genes³², we estimate there to be 60,000 exonic SNPs in this collection. The density of SNPs in exons (one SNP per 1.08 kb; Table 1) is higher than in the genome as a whole, owing to the contribution of efforts targeted to exonic regions.

We also assessed the distribution of SNPs in the genomic locus surrounding each of the RefSeq mRNAs. We assigned the RefSeq exons to their genomic locations, restricting analysis to the 2,960 RefSeq mRNAs mapping onto finished sequence. As we cannot define the extent of the noncoding (regulatory) regions of each gene, we arbitrarily defined each 'gene locus' as extending from 10 kb upstream of the start of the first exon to the end of the last exon. By this definition, 93% of gene loci contain at least one SNP, and 98% are within 5 kb of the nearest SNP; also, 59% of gene loci contained five or more SNPs, and 39% ten or more. Of 24,953

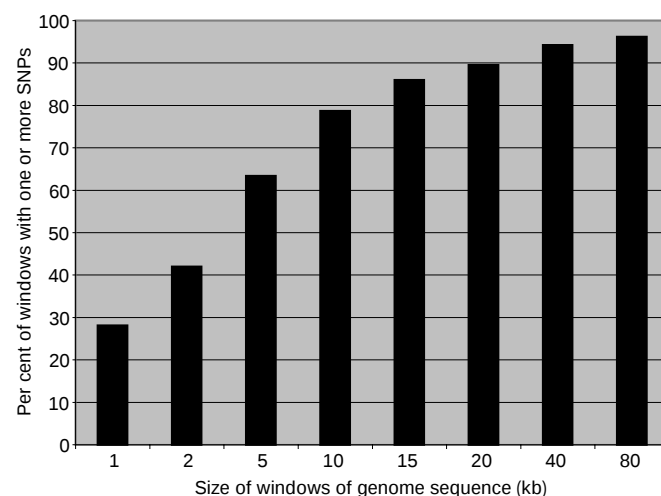


Figure 1 Distribution of SNP coverage across intervals of finished sequence. Windows of defined size (in chromosome coordinates) were examined for whether they contained one or more SNPs. Analysis was restricted to the 900 Mb of available finished sequence.

exons, 85% were within 5 kb of the nearest SNP. Thus, most exons should be close enough to at least one SNP for haplotype-based association studies, where the functional variant may be some distance from the SNPs used in the study.

The density of SNPs obtained at any given location depends upon the methods of SNP discovery contributing at each position (TSC, BAC overlap or targeted), the availability of genome sequence for SNP discovery and mapping, and the rate of nucleotide diversity. Of these, only nucleotide diversity is a fundamental characteristic of the region and population studied. To chart the landscape of human genome sequence polymorphism, we performed a genome-wide analysis of nucleotide diversity.

Analysis of nucleotide diversity

Describing the underlying pattern of nucleotide diversity required a polymorphism survey performed at high density, in a single, defined population sample, and analysed with a uniform set of tools. We reanalysed 4.5M passing sequence reads generated by TSC using genomic alignment using the NQS (see Methods). This set contained 1.2 billion aligned bases and 920,752 heterozygous positions. We measured nucleotide sequence variation using the normalized measure of heterozygosity (π), representing the likelihood that a nucleotide position will be heterozygous when compared across two chromosomes selected randomly from a population. π also estimates the population genetic parameter $\Theta = 4N_e\mu$ in a model in which sites evolve neutrally, with mutation rate μ , in a constant-sized population of effective size N_e . For the human genome, π was 7.51×10^{-4} , or one SNP for every 1,331 bp surveyed in two chromosomes drawn from the NIH diversity panel. This value agrees with smaller surveys of human genome variation^{18–20}.

We next examined the heterozygosity of individual chromosomes (Table 2). The autosomes were quite similar to one another, with 20 of 22 within 10% of the genome-wide average for autosomes (7.65×10^{-4}). Two had more extreme values: chromosome 21 ($\pi = 5.19 \times 10^{-4}$) and chromosome 15 ($\pi = 8.79 \times 10^{-4}$). Whether these observations are due to statistical fluctuations or methodological issues, or are biologically meaningful, will require investigation. The most striking difference in heterozygosity is the lower diversity of the sex chromosomes. The lower rate of polymorphism on the X chromosome may be explained by both a lower effective population

size (N_e) and lower mutation rate (μ) in $\Theta = 4N_e\mu$. Because the X chromosome is hemizygous in males, the effective population size is three-quarters of that of the autosomes. In addition, μ is higher in male than in female meiosis, with $\mu_{\text{male}}/\mu_{\text{female}} \approx 1.7/1.0$ (ref. 33). As the X chromosome undergoes male meiosis only 1/3 of the time, the overall rate of mutation in the X chromosome is expected to be 91% that of the autosomes ($\mu_X = 1.23/1.35 = 0.91$). Thus, the diversity of the X chromosome is predicted to be 69% that of the autosomes. The observed heterozygosity of the X chromosome was 4.69×10^{-4} , or 61% of the average value of the autosomes. Thus, the population genetic considerations described above could largely explain the lower heterozygosity on the X chromosome. It is possible that strong selection on the X chromosome (owing to hemizygosity in males) or other factors might partially explain this observation.

The Y chromosome has the lowest observed heterozygosity of any chromosome. It is divided into two regions: a pseudoautosomal region at either telomeric end that recombines with the X chromosome and is highly heterozygous³⁴, and the non-recombining Y (NRY). The genome assembly used for this analysis contains only the NRY, which shows very little diversity: 348 SNPs in 2,304,916 bases ($\pi = 1.51 \times 10^{-4}$). These values agree reasonably with previous estimates for NRY^{35,36}. The lower diversity of NRY is influenced by a smaller effective population size (20% that of the autosomes), counterbalanced by the higher mutation rate of male meiosis ($\mu_Y = 1.7/1.35 = 1.26 \times$ that of the autosomes). These factors predict that the Y chromosome would have a diversity 31% that of the autosomes, as compared to the observed 20%. Other influences might include selection against deleterious alleles, patterns of male dispersal³⁵ and a correlation of diversity with recombination rate¹⁹.

To look at diversity on a finer scale, we divided each chromosome into contiguous 200,000-bp bins according to the public Genome Assembly of 5 September 2000. The distribution of heterozygosity among these bins ranges from zero (12 bins, each with zero SNPs over an average of 24,720 bp examined) to 60×10^{-4} (357 SNPs in a bin surveying 58,755 bp). Although 95% of bins display nucleotide diversity values between 2.0×10^{-4} and 15.8×10^{-4} , the pattern is variable (Fig. 2a, b; see also Supplementary Information). One measure of the spread in the data is the coefficient of variation (CV), the ratio of the standard deviation (σ) to the mean (μ) of the heterozygosity π of each individual read. For the observed data, the CV ($\sigma_{\text{observed}}/\mu_{\text{observed}}$) was 1.93, considerably larger than would be expected if every base had uniform diversity, corresponding to a Poisson sampling process ($\sigma_{\text{Poisson}}/\mu_{\text{Poisson}} = 1.73$). It was expected that the observed distribution would be much more variable than a Poisson process, because both biochemical and evolutionary forces cause diversity to be nonuniform across the genome. Biological

Table 2 Nucleotide diversity by chromosome

Chromosome	Heterozygous positions	High-quality bp examined	$\pi (\times 10^{-4})$
1	71,483	92,639,616	7.72
2	81,860	111,060,861	7.37
3	61,190	81,359,748	7.52
4	59,922	74,162,156	8.08
5	56,344	77,924,663	7.23
6	53,864	72,380,717	7.44
7	52,010	68,527,550	7.59
8	44,477	57,476,056	7.74
9	41,329	50,834,047	8.13
10	43,040	52,184,561	8.25
11	47,477	56,680,783	8.38
12	38,607	51,160,578	7.55
13	35,250	43,915,606	8.03
14	35,083	47,425,180	7.40
15	27,847	31,682,199	8.79
16	22,994	27,736,356	8.29
17	21,247	27,124,496	7.83
18	24,711	30,357,102	8.14
19	11,499	15,060,544	7.64
20	22,726	31,795,754	7.15
21	26,160	50,367,158	5.19
22	17,469	20,478,378	8.53
X	23,818	50,809,568	4.69
Y	348	2,304,916	1.51
Total	920,752	1,225,448,590	7.51

Heterozygosity (π) of each chromosome. The data were filtered to remove repetitive sequences and heterozygosity calculated as described in the methods. Heterozygous positions and high-quality bases examined were counted separately for each pairwise comparison of read to genome, and then summed over each chromosome.

Table 3 Coefficients of variation for the observed data and the Poisson and coalescent models

SNPs per read	Observed	Poisson	Coalescent
0	8,796 $\pm$ 43	8,256 $\pm$ 52	8,767 $\pm$ 50
1	2,247 $\pm$ 44	3,040 $\pm$ 49	2,332 $\pm$ 46
2	668 $\pm$ 24	617 $\pm$ 24	663 $\pm$ 26
3	214 $\pm$ 14	99 $\pm$ 9	200 $\pm$ 15
4	102 $\pm$ 10	16 $\pm$ 4	66 $\pm$ 9
σ/μ	1.94 $\pm$ 0.02	1.72 $\pm$ 0.02	1.96 $\pm$ 0.03

Observed distribution of heterozygosity and comparison to expectation under Poisson and coalescent population genetic models. The autosomes were divided into 200,000-bp bins according to chromosome coordinates and one read randomly selected from each bin. This procedure was chosen to minimize the correlation in gene history of nearby regions, under the simplifying assumption that reads 200,000 bp apart and selected from unrelated individuals will have uncorrelated genealogies. Correlation of gene history does not influence the expected mean value of the CV, but does effect its variance. The random selection of reads and generation of expected distributions were repeated 100 times: presented are the mean and standard deviation of the number of reads in which 0, 1, 2, 3, or 4 SNPs were observed or predicted under each scenario. The Poisson model reports the number of such reads expected to display 0–4 SNPs under Poisson sampling of each read with a heterozygosity adjusted for length and GC content (Fig. 2c). Even in this reduced data set, the Poisson model can be rejected at $P < 10^{-99}$. The coalescent simulation³⁸ assumed a constant-sized population of effective size 10,000 and free recombination among reads. For each read, μ was scaled according to its length and GC content (Fig. 2c). Each sampled read was assigned a coalescent history from a simulated distribution and the number of SNPs predicted. The coefficient of variation of the estimate of heterozygosity is presented, with the mean and standard deviation of the 100 sampling runs shown.

factors may include rates of mutation and recombination at each locus. For example, heterozygosity is correlated with the GC content for each read (Fig. 2c), reflecting, at least in part, the high frequency of CpG to TpG mutations arising from deamination of methylated 5-methylcytosine. Population genetic forces are likely to be even more important: each locus has its own history, with samples at some loci tracing back to a recent common ancestor, and other loci describing more ancient genealogies. The time to the most recent common ancestor at a particular stretch of DNA is variable, and represents the opportunity for sequence divergence; thus, the expected pattern of heterozygosity is more heterogeneous than if every locus shared the same history^{37,38}.

To assess whether gene history would account for the observed variation in heterozygosity, we compared the observed CV to that expected under a standard coalescent population genetic model. For

each read, we adjusted μ on the basis of its per cent GC and length, and simulated genealogical histories under the assumption of a constant-sized population with $N_e = 10,000$. The CV determined under this model ($\sigma_{\text{constant-size}}/\mu_{\text{constant-size}} = 1.96$) is a close match to the observed data. To estimate standard deviations around these estimates of the CV, it was necessary to consider that tightly linked regions may display correlated histories, and thus are nonindependent. We sampled subsets of the data chosen to minimize correlation among reads (see Methods), providing estimates of the mean and standard deviation of CV for the observed and simulated data (Table 3). These results indicate that the observed pattern of genome-wide heterozygosity is broadly consistent with predictions of this standard population genetic model (for comparison, see an analysis of variation in heterozygosity in the mouse genome)³⁹. However, much work will be required to assess additional factors

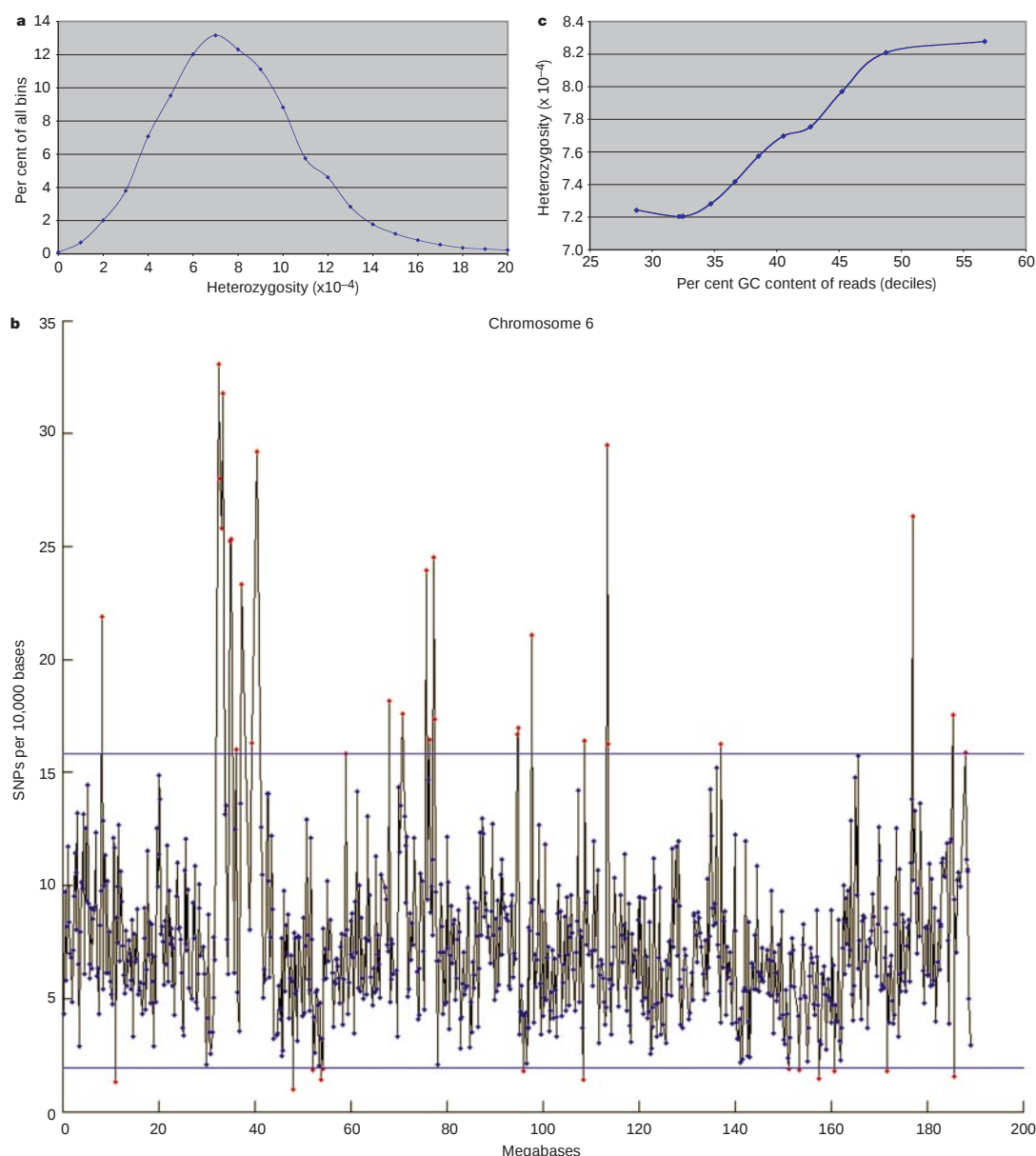


Figure 2 Distribution of heterozygosity. **a**, The genome was divided into contiguous bins of 200,000 bp based on chromosome coordinates, and the number of high-quality bases examined and heterozygosity calculated for each. A histogram was generated of the distribution of heterozygosity values across all such bins. **b**, Heterozygosity was calculated across contiguous 200,000-bp bins on Chromosome 6. The blue lines represent the values within which 95% of regions fall: 2.0×10^{-4} – 15.8×10^{-4} . Red, bins falling outside

this range. The extended region of unusually high heterozygosity centred at 34 Mb corresponds to the HLA. **c**, Correlation of nucleotide diversity with GC content of each read (autosomes only). The GC content and heterozygosity of reads from the heterozygosity analysis was calculated after sorting of reads by GC content and separation into 10 bins of equal size. Each bin contains ~150 Mb of aligned, high-quality sequence.

that could influence this distribution: biological factors such as variation in mutation and recombination rates, historical forces such as bottlenecks^{40,41}, expansions or admixture of differentiated populations, evolutionary selection, and methodological artefacts.

Regions of low diversity were more prevalent on the sex chromosomes. Whereas only 2.5% of 200,000-bp bins across the genome had $\pi < 2.0 \times 10^{-4}$, 15% of bins on the X chromosome⁴² and 89% on the Y chromosome (NRY) had these levels of diversity. Regions of low diversity may be explained by the smaller effective population size of the sex chromosomes and the variable underlying distribution of heterozygosity. Strong selection acting on the sex chromosomes in males might also have a role, but this hypothesis requires further testing. Regions of high heterozygosity were also observed. One was found on chromosome 6 (Fig. 2b, centred on 34 Mb), and was confirmed to represent the HLA locus, which has high nucleotide diversity owing to balancing selection⁴³. Other regions of varying size were observed on this and other chromosomes (Fig. 2c and Supplementary Information). Some of these highly diverse regions might have also experienced balancing selection, but there are other possible explanations: for example, sampling fluctuations of the coalescent distribution, regions with high rates of mutation and/or recombination, unrecognized duplications in the human genome and sequencing of a rare haplotype by the HGP (to which the TSC reads were compared).

Given the unfinished state of publicly available sequence data and genome assembly, it will be important to reevaluate these estimates as more complete genome sequence becomes available.

Implications for medical and population genetics

We describe a map of publicly available SNPs (as of November 2000), fully integrated with the sequence, physical and genetic maps of the human genome. We anticipate immediate application to studies of human population genetics, candidate-gene studies for disease association, and eventually unbiased, genome-wide association scans. First, the map provides an unprecedented tool for studying the character of human sequence variation. We use these data to describe the first genome-wide view of how human DNA sequence varies in the population, and the public availability of these data should fuel future research into biological and population genetic influences on human genetic diversity.

Second, insights into human evolutionary history will be obtained by using SNPs from the map to characterize haplotype diversity throughout the genome. Human haplotype structure remains largely unexplored, and this map makes it possible to define the extent and variation of haplotype identity, the number and frequencies of common haplotypes, and their distribution among and within existing ethnic groups.

Most practically, where a gene has been implicated in causing disease (by chromosomal position relative to linkage peaks, known biological function or expression pattern), it is desirable exhaustively to survey allelic variation for any association to disease. Using the SNP map, it should be possible to evaluate the extent to which common haplotypes contribute to disease risk. As the speed and efficiency of SNP genotyping increases, such studies will fuel increasingly comprehensive tests of the hypothesis that common variants contribute significantly to the risk of common diseases. To the extent that such studies are successful, they should profoundly affect our understanding of disease, methods of diagnosis, and ultimately the development of new and more effective therapies. □

Methods

SNP identification

Candidate SNPs were identified by detection of high-confidence base differences in aligned sequences. For TSC, sequence reads were filtered to exclude low quality reads and those containing predominantly known repetitive sequence. Sequences were aligned to each other using the reduced representation shotgun (RRS) method, and by genomic alignment (GA) as described^{18,22}. For GA of TSC data, reads were compared to available

large-insert clones (finished and draft with available PHRAP quality scores) in Genbank. For the analysis of clone overlaps, all available finished and unfinished genomic sequence accessions were aligned. Two methods were used to detect SNPs. The NQS relies upon the sequence trace quality surrounding the SNP base to increase base-calling confidence^{18,22}; most data discovered using the NQS was processed using SsahaSNP, an ultrafast, hash-based implementation of the algorithm (Z.N., A. Cox and J.C.M., manuscript in preparation). The second method calculates confidence scores on the basis of a Bayesian analysis of confidence scores²⁴. A variety of methods were used to find SNPs in expressed sequence tag (EST) overlaps^{24,25,27} and for targeted resequencing; details of the remaining SNPs can be found in the individual dbSNP entries (www.ncbi.nlm.nih.gov/SNP/).

Mapping of SNPs and features

MEGABLAST⁴⁴ was used to align TSC SNP flanking sequences to the genomic sequence accessions. A SNP was considered mapped if a high-quality match (99% identity or greater) was found across the available flanking sequence of no less than 270 bp. SNPs that matched more than three accessions with identity > 98% were judged to be possible repetitive regions and set aside. SNP coordinates were generated relative to the OO18 build of the genome assembly (5 September 2000) and the OO15 build (15 July 2000), using the AGP format files provided by D. Haussler (<http://genome.ucsc.edu>).

The NCBI RefSeq mRNA transcripts³¹ were aligned to the Genome Assembly using the NCBI SPIDEY alignment tool. Alignment required >97% sequence similarity between mRNA and genome sequence; alignments were refined by taking into account the donor/acceptor sites. In cases where CDS annotations were available in the GenBank record, exons of the CDS were aligned within the confines of the mRNA alignment. Regions of known human repeats were annotated directly using RepeatMasker (A. Smit, unpublished).

Nucleotide diversity analysis

To characterize nucleotide diversity, we required a data set in which all data could be analysed both for the number of high-quality bases meeting quality standards for SNP detection, and for the number of SNPs. To ensure homogeneity of analysis, we performed a single analysis of 4.5 million high-quality TSC reads from the Sanger Centre, Washington University in St. Louis and the Whitehead Center for Genome Research. The GC content of these reads was 41%, the same as the genome as a whole³², and the distribution of read GC content across deciles of the genome (sorted by GC content) was within 10% of the expected value for all bins. The read coverage was well distributed: 88% of contiguous 200,000-bp windows contained over 10,000 aligned bases (5%) surveyed for SNPs (see below). Using a single analytic tool (SsahaSNP, an implementation of the NQS; Z.N., A. Cox and J.C.M., in preparation), these reads were aligned to the available genome sequence (finished and draft with quality scores) and the number of high-quality bases (meeting NQS) and SNPs counted. We limited the analysis to SNPs found by genomic alignment so that the cluster depth of each comparison would be exactly two chromosomes. We precisely measured the target size for SNP discovery by counting the number of positions meeting the NQS. This is desirable because alignments contain positions of both high and low quality, but only those meeting the NQS are candidates for SNP discovery. Where a single TSC read aligned to multiple (overlapping) BACs from the HGP, we averaged the number of SNPs and aligned bp for all pairwise alignments of that read; this weighted evenly those reads mapping to a single BAC and those aligning to a region of overlap. Reads representing repeat loci were excluded using validated criteria^{18,22}: alignments of reads to genome were excluded if they were less than 99% identical. The genome was then divided into contiguous bins of 200,000 bp (based on chromosome-relative coordinates). Individual reads were filtered for repeats: any that aligned to more than one bin in the genome assembly were rejected. Finally, heterozygous positions and bases meeting the NQS were counted. As a final filter for regions containing a high proportion of repeats, we reject any bin for which more than 10% of the reads mapping to that bin also mapped to another chromosome. Finally, to avoid statistical fluctuation due to inadequate sampling, we examined only the 88% of bins in which at least 10,000 aligned bases met the NQS and thus could be examined for SNPs.

Coalescent modelling was performed by simulation³⁸, and assumed a constant-sized population of 10,000 individuals and a mutation rate adjusted for each read on the basis of its GC content (Fig. 2c). To assess the standard deviation around this estimate, the simulation was repeated 100 times. For the observed data, calculating a standard deviation around the CV is difficult owing to the correlation of gene history for closely linked sites. In expectation, this correlation should not alter the mean of the observed coefficient of variation, but does influence its variance. To estimate the variance around the CV for the observed data, we selected 100 reduced data sets, each containing one randomly chosen read from each 200,000-bp bin along the autosomes. In using this approach, we assume that these reads, 200,000 bp apart and sampled from unrelated individuals, have independent genealogies. This random sampling procedure was repeated 100 times to estimate the mean and variance of the observed CV.

The data for the heterozygosity analysis, including the coordinates of each bin, the number of bases examined and number of SNPs identified, is available as Supplementary Information.

Received 28 November; accepted 27 December 2000.

- Collins, F. S. Of needles and haystacks: finding human disease genes by positional cloning. *Clin. Res.* **39**, 615–623 (1991).
- Collins, F. S., Guyer, M. S. & Charkravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
- Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
- Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).

5. Li, W. H. & Sadler, L. A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
6. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes [published erratum appears in *Nature Genet.* **23**, 373 (1999)]. *Nature Genet.* **22**, 231–238 (1999).
7. Cambien, F. *et al.* Sequence diversity in 36 candidate genes for cardiovascular disorders. *Am. J. Hum. Genet.* **65**, 183–191 (1999).
8. Fullerton, S. M. *et al.* Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am. J. Hum. Genet.* **67**, 881–900 (2000).
9. Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
10. Nickerson, D. A. *et al.* DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.* **19**, 233–240 (1998).
11. Rieder, M. J., Taylor, S. L., Clark, A. G. & Nickerson, D. A. Sequence variation in the human angiotensin converting enzyme. *Nature Genet.* **22**, 59–62 (1999).
12. Templeton, A. R., Weiss, K. M., Nickerson, D. A., Boerwinkle, E. & Sing, C. F. Cladistic structure within the human lipoprotein lipase gene and its implications for phenotypic association studies. *Genetics* **156**, 1259–1275 (2000).
13. Eaves, I. A. *et al.* The genetically isolated populations of Finland and sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nature Genet.* **25**, 320–323 (2000).
14. Taillon-Miller, P. *et al.* Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genet.* **25**, 324–328 (2000).
15. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
16. Collins, A., Lonjou, C. & Morton, N. E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96**, 15173–15177 (1999).
17. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* (submitted).
18. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
19. Nachman, M. W., Bauer, V. L., Crowell, S. L. & Aquadro, C. F. DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**, 1133–1141 (1998).
20. Wang, D. G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
21. Jorde, L. B. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**, 1435–1444 (2000).
22. Mullikin, J. C. *et al.* An SNP map of human chromosome 22. *Nature* **407**, 516–520 (2000).
23. Collins, F. S., Brooks, L. D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation [published erratum appears in *Genome Res.* **9**, 210 (1999)]. *Genome Res.* **8**, 1229–1231 (1998).
24. Marth, G. T. *et al.* A general approach to single-nucleotide polymorphism discovery. *Nature Genet.* **23**, 452–456 (1999).
25. Buetow, K. H., Edmonson, M. N. & Cassidy, A. B. Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genet.* **21**, 323–325 (1999).
26. Gu, Z., Hillier, L. & Kwok, P. Y. Single nucleotide polymorphism hunting in cyberspace. *Hum. Mutat.* **12**, 221–225 (1998).
27. Irizarry, K. *et al.* Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nature Genet.* **26**, 233–236 (2000).
28. Picoult-Newberg, L. *et al.* Mining SNPs from EST databases. *Genome Res.* **9**, 167–174 (1999).
29. Marth, G. T. *et al.* Single nucleotide polymorphisms in the public database: how useful are they? *Nature Genet.* (submitted).
30. Yang, Z. *et al.* Sampling SNPs. *Nature Genet.* **26**, 13–14 (2000).
31. Pruitt, K. D., Katz, K. S., Sicotte, H. & Maglott, D. R. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* **16**, 44–47 (2000).
32. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
33. Bohossian, H. B., Skaletsky, H. & Page, D. C. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature* **406**, 622–625 (2000).
34. Cooke, H. J., Brown, W. R. & Rappold, G. A. Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. *Nature* **317**, 687–692 (1985).
35. Shen, P. *et al.* Population genetic implications from sequence variation in four Y chromosome genes. *Proc. Natl Acad. Sci. USA* **97**, 7354–7359 (2000).
36. Underhill, P. A. *et al.* Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.* **7**, 996–1005 (1997).
37. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
38. Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. & Antonovics, J.) 1–44 (Oxford Univ. Press, Oxford, 1991).
39. Lindblad-Toh, K. *et al.* Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genet.* **24**, 381–386 (2000).
40. Kimmel, M. *et al.* Signatures of population expansion in microsatellite repeat data. *Genetics* **148**, 1921–1930 (1998).
41. Reich, D. E. & Goldstein, D. B. Genetic evidence for a Paleolithic human population expansion in Africa [published erratum appears in *Proc. Natl Acad. Sci. USA* **95**, 11026 (1998)]. *Proc. Natl Acad. Sci. USA* **95**, 8119–8123 (1998).
42. Miller, R. D., Taillon-Miller, P. & Kwok, P. Y. Regions of low single-nucleotide polymorphism (SNP) incidence in human and orangutan Xq: deserts and recent coalescences. *Genomics* (in the press).
43. Horton, R. *et al.* Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. *J. Mol. Biol.* **282**, 71–97 (1998).
44. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214 (2000).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

The SNP Consortium, the Wellcome Trust and the National Human Genome Research Institute funded SNP discovery and data management at Cold Spring Harbor Laboratories, The Sanger Centre, Washington University in St. Louis, and the Whitehead/MIT Center for Genome Research. Work in P.Y.K.'s laboratory is supported in part by grants from the SNP Consortium and the National Human Genome Research Institute. P.Y.K. thanks Q. Li, M. Minton, R. Donaldson and S. Duan for technical assistance. D.M.A. was supported during a phase of this work under a Postdoctoral Fellowship for Physicians from the Howard Hughes Medical Institute. For full list of contributors to TSC programme, see www.snp.cshl.org.

Correspondence and requests for materials should be addressed to D.A. (e-mail: altshul@genome.wi.mit.edu) or D.B. (e-mail: drb@sanger.ac.uk).

* **The International SNP Map Working Group** (contributing institutions are listed alphabetically).

Cold Spring Harbor Laboratories: Ravi Sachidanandam¹, David Weissman¹, Steven C. Schmidt¹, Jerzy M. Kakol¹ & Lincoln D. Stein¹

National Center for Biotechnology Information: Gabor Marth² & Steve Sherry²

The Sanger Centre: James C. Mullikin³, Beverley J. Mortimore³, David L. Willey³, Sarah E. Hunt³, Charlotte G. Cole³, Penny C. Coggill³, Catherine M. Rice³, Zemin Ning³, Jane Rogers³, David R. Bentley³

Washington University in St. Louis: Pui-Yan Kwok⁴, Elaine R. Mardis⁴, Raymond T. Yeh⁴, Brian Schultz⁴, Lisa Cook⁴, Ruth Davenport⁴, Michael Dante⁴, Lucinda Fulton⁴, LaDeana Hillier⁴,

Robert H. Waterston⁴ & John D. McPherson⁴

Whitehead/MIT Center for Genome Research: Brian Gilman⁵, Stephen Schaffner⁵, William J. Van Etten^{5,6}, David Reich⁵, John Higgins⁵, Mark J. Daly⁵, Brendan Blumenstiel⁵, Jennifer Baldwin⁵, Nicole Stange-Thomann⁵, Michael C. Zody⁵, Lauren Linton⁵, Eric S. Lander^{5,7} & David Altshuler^{5,8}

1, Cold Spring Harbor, New York 11724, USA; 2, Building 38A, 8600 Rockville Pike, Bethesda, Maryland 20894, USA; 3, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK; 4, 660 S. Euclid Ave, St. Louis, Missouri 63110, USA; 5, 9 Cambridge Center, Cambridge, Massachusetts 02139, USA; 6, Present address: Blackstone Technology Group, Boston, Massachusetts 02110, USA; 7, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA; 8, Departments of Genetics and Medicine, Harvard Medical School; Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.

A physical map of the human genome

The International Human Genome Mapping Consortium*

* A partial list of authors appears at the end of this paper. A full list is available as Supplementary Information.

The human genome is by far the largest genome to be sequenced, and its size and complexity present many challenges for sequence assembly. The International Human Genome Sequencing Consortium constructed a map of the whole genome to enable the selection of clones for sequencing and for the accurate assembly of the genome sequence. Here we report the construction of the whole-genome bacterial artificial chromosome (BAC) map and its integration with previous landmark maps and information from mapping efforts focused on specific chromosomal regions. We also describe the integration of sequence data with the map.

The International Human Genome Sequencing Consortium (IHGSC) used a hierarchical mapping and sequencing strategy to construct the working draft of the human genome. This clone-based approach involves generating an overlapping series of clones that covers the entire genome. Each clone is 'fingerprinted' on the basis of the pattern of fragments generated by restriction enzyme digestion^{1,2}. Clones are then selected for shotgun sequencing and the whole genome sequence is reconstructed by map-guided assembly of overlapping clone sequences³.

The availability of the whole-genome clone-based map assisted the sequencing of the human genome in many respects. The fingerprinted BAC map made it possible to select clones for sequencing that would ensure comprehensive coverage of the genome and reduce sequencing redundancy. In addition, the challenge of sequence assembly was minimized by restricting random shotgun sequencing to individual clones. Furthermore, the clone-based map also enabled the identification of large segments of the genome that are repeated, thereby simplifying the assembly. Many IHGSC centres had developed chromosomal maps and resources that were not integrated, so it was essential to have a unifying genome map to enable localization of clones, with respect to previously sequenced clones, before they were sequenced. The accurate fingerprinting and sizing of each clone enabled us to verify the accuracy of shotgun sequence⁴ assembly of each clone.

The human genome presented unique challenges for the development of a clone-based physical map. Its size of 3.2 gigabases (Gb), which is 25 times as large as any previously mapped genome, meant that proportionately more clones had to be analysed. Its greater complexity also made it more difficult to distinguish true overlaps, which was further complicated by the repeat-rich nature of the genome. Early efforts to construct clone-based regional and even chromosomal physical maps of the human genome using cosmid libraries derived from isolated human chromosomes met with limited success^{5,6}. By contrast, maps based on sequence-tagged site (STS) landmarks provided greater coverage of the genome⁷⁻⁹, as did genetic maps based on variations in simple sequence repeats in STS landmarks^{10,11}. The development of P1-artificial chromosome (PAC)¹² and bacterial artificial chromosome (BAC)¹³ cloning systems was pivotal to the success of the whole-genome map. They provided larger inserts, more stable clones and better coverage of the genome.

Clone-based maps similar to that described here have been important in the sequencing of most large genomes, including those of *Saccharomyces cerevisiae*¹, *Caenorhabditis elegans*² and *Arabidopsis thaliana*¹⁴. A clone-based map also contributed to the sequencing of the *Drosophila melanogaster* genome^{15,16} and a combined mapping and sequencing strategy is being applied to the mouse genome^{17,18}. This work illustrates the benefit of using the clone-based map in the assembly of the human genome sequence.

Construction of the whole-genome BAC map

The pilot phase of the sequencing project began in 1995, at which time efforts were renewed to develop clone-based maps covering specific regions of the genome. To construct these regional maps, we screened PAC and BAC clones for STS markers, fingerprinted the positive clones, integrated them into the existing maps, and selected the largest, intact clones with minimal overlap for sequencing.

To keep pace with the ramping up of the sequencing effort in 1998, the ongoing efforts to construct the whole-genome BAC map were increased approximately tenfold. The whole-genome BAC map was constructed in several steps. First we collected fingerprint data for a large sample of random clones from a genome-wide BAC library. We then assembled the BAC map, first by using the fingerprint data to cluster highly related clones automatically, then by further refining them manually, and last by merging contigs with related clones at their ends. Finally, in parallel with construction of the BAC map, we mapped the chromosomal positions of individual clones on the basis of landmarks from existing landmark maps.

Fingerprinting the BAC clones

In October 1998, we began fingerprinting 300,000 BACs from the RPCI-11 library¹⁹ (<http://www.chori.org/bacpac/>). Redundancy of sampling was vital to achieve high continuity in the final map¹⁴. Assuming an average BAC insert size of 150,000 base pairs (bp) and a genome size of 3.2 Gb, this level of fingerprinting would provide roughly 15-fold coverage of the genome. The library was derived from male DNA, providing full coverage of all 24 human chromosomes but with half as much coverage of the sex chromosomes as of the autosomes. Our experience with the library found it to be of high quality with uniformly large-insert clones, few non-recombinant clones and little cross-contamination of source plates. The RPCI-11 library was one of the first libraries to meet the informed consent criteria in accordance with the NHGRI policy for the Use of Human Subjects in Large Scale Sequencing (http://www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/Human_subjects.html).

To meet the goal of fingerprinting 300,000 BAC clones in one year, we devised a tandem 121-lane agarose gel format, allowing the simultaneous electrophoresis of 50 standard 'marker' DNA lanes and of 192 BAC restriction digests (Fig. 1), thereby reducing the number of gels, without loss of restriction fragment size accuracy or fidelity of clone tracking (see Supplementary Information). With these and other improvements in the fingerprinting technology and resources, we increased throughput tenfold to process more than 20,000 fingerprints (which equates to approximately onefold clone coverage of the human genome) each week. We also sampled clones from the RPCI-13 and CT-C/D1 BAC libraries, which were constructed using a different restriction enzyme (Table 1). This provided differential sampling of the genome, given the different distribution of the restriction enzyme sites within the genome. In

addition, the RPCI-13 library is derived from female DNA, which improves the representation of the X chromosome in the whole-genome BAC map.

Assembling the BAC map

By the end of 1999, with the fingerprint data on the BAC clones entered into an FPC database^{20–23} (<http://www.sanger.ac.uk/Software/fpc/>), we were ready to construct the initial fingerprint assembly that would form the basis for further work on the map. We experimented with various strategies for automated assembly that would be as complete and as consistent as possible (see Supplementary Information).

First, we edited the fingerprint data itself. In early tests of assembly, we found that the variability in the mobility of small fragments (< 600 bp) led to artefactually low estimates of overlaps between clones. We therefore removed fragments smaller than 600 bp before assembly. Similarly, variability in estimating band numbers in 'multiplets' (instances where more than one fragment is located at nearly the same position on the gel) also caused problems. To reduce the variability between the number of bands called in these multiplet situations and thus increase the reliability with which related clones are correctly overlapped, these fragments were collapsed to only a single band in the resulting fingerprint. This 'sanitizing' process resulted in clusters of increased reliability.

Second, we evaluated the impact of varying the threshold for the 'overlap statistic', which is a measure of clone similarity, and the tolerance for accepting two bands from different clones as the same. We compared the clusters obtained for consistency with known regions and with other mapping data for the fingerprinted clones (primarily radiation hybrid chromosomal localization data from the Stanford Human Genome Center (SHGC)). The parameters finally used (overlap statistic of 3×10^{-12} or about 75% clone overlap and 0.7 mm tolerance) balanced the total number of clusters (which decreased with less stringent parameters) and the number of chimaeric clusters (which decreased with more stringent parameters). The automated assembly of 283,287 BAC clones resulted in 7,133 clusters containing 93% of all fingerprints in the database (Table 2). The remaining unincorporated clones (singletons) were excluded, as they contained too few bands to be included by automated assembly under these conditions or simply had no closely related clones. These latter clones included artefacts such as clones that had rearranged or had poor quality data, as well as rare clones representing poorly sampled portions of the genome.

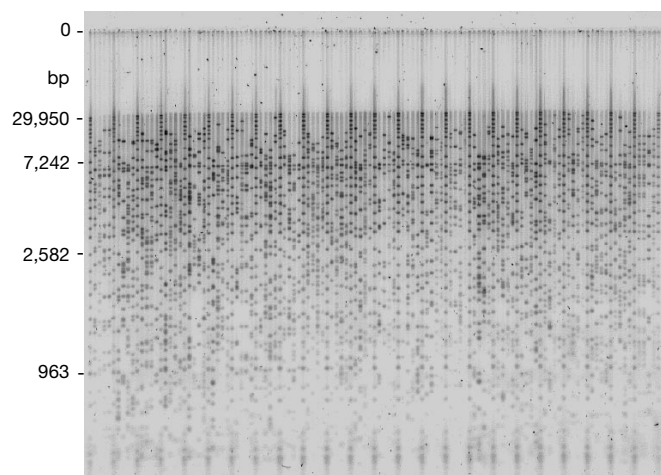


Figure 1 Example of the improved high-throughput fingerprint gel. BAC DNAs are digested with *HindIII* and visualized on a SYBR-green-stained 1% agarose gel. Every fifth lane contains a mixture of marker DNAs; the sizes of selected marker fragments are indicated. 0, origin of fragment migration.

Table 1 Sources of clones used

Library	Clones in current whole-genome map	Type	Vector	Enzyme	Average insert size (kb)
RPCI-4, -5*	568	PAC	pCYPAC2	<i>Mbol</i>	116
RPCI-11*	272,027	BAC	pBACe3.6	<i>EcoRI</i>	174
			pTARBAC1†	<i>Mbol</i>	196
RPCI-13*	59,051	BAC	pBACe3.6	<i>Mbol</i> or <i>DpnII</i>	149
CT-A, -B‡	228	BAC	pBeloBAC11	<i>HindIII</i>	120
CT-C, -D1‡	52,725	BAC	pBeloBAC11	<i>HindIII</i>	125
CT-D2‡	559	BAC	pBeloBAC11	<i>EcoRI</i>	190
Other§	10,231				

* <http://www.chori.org/bacpac/>

† RPCI-11, segment 5

‡ http://informa.bio.caltech.edu/Bac_info.html

§ Various clones from multiple libraries sent by collaborating centres.

As fingerprints from new clones were added after the initial assembly, there was a disproportionate increase in the number of singletons (Table 2). These new data were only incorporated into existing clusters or contigs if they added needed depth or helped to join contigs. We noted a further increase in singletons as new libraries were sampled (particularly from the RPCI-13 and CT-C/D1 libraries). One possible explanation is that these new libraries encompass regions of the genome not represented in the initial RPCI-11 library.

Most clones (97.5%) in the current whole-genome BAC map are derived from RPCI-11 (272,027/69.2%), RPCI-13 (59,051/14.9%) and CT-C/D1 (52,725/13.3%) (Table 1). Although only about two-thirds of the fingerprint data are derived from DNA from a single individual, we did not experience any problems in assembly arising from polymorphisms between the individuals from whom the DNA was obtained.

Achieving map continuity

The goals of the manual editing were to refine the ordering of the clones within clusters to create contigs, to disassemble larger chimaeric contigs (representing clusters of two or more sets of non-overlapping clones) and to join contigs. This process involved first editing the fingerprint assemblies (using the tools encapsulated in FPC) to ensure that every clone within a contig was properly situated with respect to its most highly related neighbours, defined by fingerprint similarity¹⁴ (see Supplementary Information). About 600 chimaeric clusters were identified and disassembled. To identify potential joins, we then used clones at the extreme ends of each contig to query the FPC database at a lower required fingerprint overlap stringency (overlap statistic of 1×10^{-8} or about 50% clone overlap) than was used during initial assembly. Joins were incorporated into the map if the fingerprinting data was logically consistent with the proposed map order (Fig. 2).

The most notable effect of the intensive editing was the greater than fivefold reduction in total contigs, from a high of 7,700 contigs after chimaeric contigs had been disassembled, to 1,246 by the 7 October 2000 data freeze of the draft genome sequence³ (Table 2). The longest contig in this set encompasses more than 60 Mb of draft genome sequence and the mean contig size is estimated to be

Table 2 Status of FPC database after automated assembly and manual editing

	Automated assembly	Manually edited database
Date	December 1999	September 2000
BAC clones in FPC	283,287	372,264
Number of contigs	7,133	1,447
Clones in contigs	264,555	295,828
Number of singletons*	18,732	76,436
Contigs containing:		
>25 clones	3,012	912
9–25 clones	1,844	260
3–9 clones	1,957	204
2 clones	887	71

* Clones not incorporated into any contig; see text.

database. The combined ePCR and hybridized data sets contained 69,507 markers, including 1,659 polymorphic markers from the Génethon genetic map. We primarily used GeneMap'99 for further anchoring and ordering of contigs, as it has a substantial marker set (> 50,000), is well integrated with the Génethon genetic map and provides local ordering at < 1 Mb resolution. Once sequenced clones could be reliably associated with the fingerprinted clones³, we could use the marker content of sequenced clones determined by ePCR to order and orient contigs more reliably. We used markers found on any of six maps (Génethon genetic map, Marshfield genetic map¹¹, WIBR YAC STS-content (http://carbon.wi.mit.edu:8000/cgi-bin/contig/phys_map), GeneMap'99, SHGC G3 radiation hybrid map (<http://www-shgc.stanford.edu/Mapping/rh/index.html>) and NCBI framework map (<http://www.ncbi.nlm.nih.gov/genome/guide/>)) to orient contigs with respect to the majority consensus of all maps examined.

Integration of specific mapping efforts

We integrated regional map data into the whole-genome BAC map from other genome centres (see list at <http://www.nhgri.nih/>), which enriched the map and helped in the selection of clones for sequencing as it minimized redundancy and improved coverage. The regional mapping data included those for chromosomes 12 (ref. 30), 14 (ref. 31) and Y (ref. 32), and 1, 6, 9, 10, 13, 20, 22 and X (ref. 33). We also integrated mapping data for chromosome 19 (Lawrence Livermore National Laboratory, http://www-bio.llnl.gov/bbrp/genome/html/chrom_map.html) and a 20-Mb segment of chromosome 15 (University of Washington). Telomeric contigs were identified and positioned where possible, as described elsewhere in this issue³⁴.

Some mapping efforts employed clone resources other than the RPCI-11 BAC library. In these cases, clones were sent by these centres, fingerprinted at the Washington University Genome Sequencing Center (WUGSC) and incorporated into the whole-genome BAC map and FPC database. These clones included those from regions of chromosomes 5 (J. Cheng), 8 (A. Rosenthal and N. Shimizu³⁵), 11 (Y. Sakaki) and 17 (J. Ramser). In addition, we used computer-generated restriction digests, or *in silico* digests, of sequences in GenBank to incorporate these clones into the whole-genome BAC map.

Table 3 Chromosomal assignment of contigs

Chromosome	Number of contigs	Estimated size (Mb)*
1	119	263
2	54	255
3	77	214
4	42	203
5	51	194
6	21	183
7	29	171
8	46	155
9	25	145
10	23	144
11	36	144
12	30	143
13	15	114
14	21	109
15	19	106
16	64	98
17	50	92
18	28	85
19	59	67
20	10	72
21	17	50
22	10	56
X	163	164
Y	8	59
UL†	229	—
Total	1,246	3,286

* Ref. 45

† UL consists of clone contigs that could not be reliably placed on a chromosome.

Accuracy of chromosomal positions

As an independent assessment of the accuracy of assigning a chromosomal position to contigs, we used aliquots of the BAC DNA from 96 fingerprinted clones (RPCI-11, clones 512M01–512O24) as FISH probes to metaphase chromosomes (see Methods). Of the 96 BACs examined, 87 were successfully assigned to a single chromosome band. The remaining clones either failed to label (six) or were associated with multiple chromosome bands (three). The chromosomal localization of 82 (94%) of the mapped BACs agreed unambiguously with the derived chromosomal assignment, based on STS content, of the contig into which the corresponding fingerprint had assembled. A single BAC mapped to one of the two positions that were equally well supported by the marker content of its associated contig. The remaining four BACs were associated with fingerprints in contigs that had no mapped marker content and thus were not previously localized. In summary, the FISH mapping data did not conflict with any of the chromosomal assignments of the contigs examined.

In addition, we selected a minimal tiling path of eight clones from a random contig. DNA remaining from the fingerprinting of these clones was used for FISH mapping. All eight clones co-localized to chromosomal segment 8q21.1. This was consistent with other FISH data (B. Trask; 8q21.1), radiation hybrid data (D. R. Cox; chromosome 8) and ePCR of 12 markers mapping to chromosome 8 (Fig. 2).

Accuracy of clone order

The integration of independent map data and the emerging sequence information enabled us to monitor the fidelity of the developing map. We regularly checked that the predicted clone order was reflected in the overlaps of the sequenced clones. The ongoing assignment of chromosomally positioned markers to clones and contigs provided a useful check for possible false joins between unrelated contigs. These checks for clone order and contig fidelity were carried out much more extensively once the draft genome sequence was assembled and additional marker data incorporated. Overall, local clone order agreed with the overlaps demonstrated by sequence.

Comparison of the chromosome 12 STS-content BAC map³⁰ with the fingerprint BAC map of the same chromosome provided an important test of the accuracy of clone ordering. The two maps were derived independently, but used the same RPCI-11 library. The maps are consistent in clone ordering and provide complementary resources: the chromosome 12 STS-content BAC map provides more accurate contig anchoring and orientation, and our map provides more depth of clone coverage. Furthermore, these maps, while sharing some gaps, largely closed gaps for each other, underscoring the benefit of the complementary mapping strategies. After integration, the resulting map consisted of four contigs on the short arm and 34 on the long arm, and this has been further reduced to 20 contigs by continued gap closure methods³⁰.

Duplications and repeats

Two problematic aspects of the genome still need to be resolved: large (> 150 kilobase (kb)) recently duplicated segments and smaller tandemly repeated sequences extending for > 100 kb. Analysis of the total clone population shows that about 1% of clones have unusually high numbers of closely related clones (>75% shared bands), indicative of large repeated sequences. In some cases, minor differences in band patterns have allowed some complex repeats to be tentatively teased apart, but many of these have yet to be investigated in detail at the sequence level (an exception is the Y chromosome³²). In other cases, only more complete and finished sequence will clarify mapping data for these regions.

The presence of extensive smaller tandemly repeated sequences (which sometimes are not even successfully cloned) results in clones that resemble small insert and badly deleted clones, which we avoided including in the map. However, unlike the small and

deleted clones, tandem repeats are present in multiple independent clones that display a similar fingerprint pattern. Sequence analysis of some of these repeat sequences shows that they are related to centromeric and ribosomal-repeat-related repeats, among others.

Gaps in the map

The remaining gaps, currently fewer than 1,000, are likely to stem from a variety of causes. There may be overlaps between end clones that are too small to be detected by fingerprints and which will only be recognized once the end clones are sequenced. Gaps can also arise because of misassemblies in the map, particularly where a duplicated segment is inappropriately designated to represent just one region. Some gaps may be detected from analysis of other BAC DNA libraries constructed using different restriction enzymes. Other gaps may arise simply because clones are not recovered at sufficient frequencies in BAC or PAC large insert libraries—clones spanning these gaps could potentially be detected in YAC libraries or might need to be recovered using special approaches.

Coverage of the genome

To estimate the fraction of the genome that was represented in the whole-genome map, we analysed chromosomes 21 and 22. Using *in silico* digest methods, we estimated the coverage of the fingerprint map encompassing finished chromosomes 21 (ref. 36) and 22 (ref. 37). Simulated 175-kb clones were created and digested *in silico* from the contiguous sequences for these chromosomes; each clone overlapped by 40%. We compared these digested simulated clones against the FPC database at high fingerprint overlap stringency. For chromosome 21, 316 simulated clones were created, of which 315 had at least 15 *Hind*III restriction fragments; clones containing fewer bands are difficult to compare. Of the 315 simulated fingerprints, 309 (98%) matched a related clone in the whole-genome BAC fingerprint FPC database. Similarly, for chromosome 22, 308 simulated clones were created. Of those, 303 had more than 15 *Hind*III restriction fragments and, when compared to the entire FPC fingerprint database, 297 (98%) found a related clone. This analysis was repeated using a 210-kb *in silico* clone size with 90% overlap, with similar results. Each of these chromosomes has four sequence gaps that are estimated to encompass 1.6% of the chromosome; therefore, the confirmed clone coverage of the euchromatic region of these chromosomes is around 96%. Collectively, these chromosomes represent approximately 3% of the genome. It is probably reasonable to extrapolate that this level of clone coverage will be found throughout the genome.

Clone selection for sequencing

The whole-genome BAC map was, and continues to be, used to select clones for sequencing. We devised algorithms for automatic high-throughput selection of BACs, which specifically chooses clones from contigs lacking sequenced clones ('seed clones') and clones that extend from already selected clones (see Supplementary Information). We took several issues into account when developing these programs. First, we had to devise methods compatible with a constantly and rapidly evolving map, which had considerable new information added to it each week. Second, we had to avoid clones representing genomic regions already sequenced from libraries other than RPCI-11. Third, we wished to select only clones not deleted or otherwise rearranged and thus faithfully represent the underlying genome.

The fingerprint map was used initially to identify nonredundant seed clones for sequencing when only a small portion of the RPCI-11 clones had been fingerprinted. As described above, we used all available forms of mapping data to localize the clones, and thus the contig. Once an appropriate contig was identified, the program looked for the largest clone (smaller than 225 kb to avoid artefacts) in the contig. The program also checked the fidelity of the clone by comparing its bands against other clones. We avoided end clones, as

they inevitably had bands that could not be confirmed. In addition, a clone registry was developed (NCBI) to track clones selected for sequencing by any centre, and contigs with these clones were also avoided, as were contigs containing clones with similarities to other clones in GenBank as detected by *in silico* digest.

The next step in automated clone selection was to extend progressively from the seed clones using tools to search for appropriately overlapping clones. Neighbouring clones were evaluated using the overlap statistic to provide a tentative clone order. Clones within a specified range of overlap statistic were evaluated for the total size of shared bands. The amount of acceptable overlap was also specified (typically 25%). Any candidate in turn was evaluated against an intermediately positioned clone to ensure that the overlap was genuine and was compared to existing data using the clone registry and *in silico* digests to avoid redundancy.

These automated tools were used until late January 2000, when the manually evaluated contigs became available, allowing the selection of minimal tiling paths based on these clone orders. In the course of generating the working draft, more than 10,000 BAC clones were selected for the sequencing pipelines at the WUGSC, Whitehead Institute for Biomedical Research and the Stanford Genome and Technology Center using these tools and the evolving whole-genome BAC map. A check of 518 overlaps between finished clones selected both manually and through the automated methods at WUGSC shows that they have an average overlap of 47.5 kb with their neighbours, or about 28%: this is an acceptable degree of overlap, given the relatively dense seeding that occurred, and the importance placed on achieving coverage.

Sequence map of the human genome

Although the whole-genome BAC map was constructed primarily to exploit the coverage of high-redundancy BAC libraries for use in sequencing the human genome, it has served to integrate the sequences in GenBank³⁸ with the physical map. This was needed to guide the long-range assembly of the working draft sequence and to identify all remaining gaps in this sequence map so that spanning clones could be selected. By using *in silico* digests to generate fragment size information that could be compared to the fingerprints in the FPC database, virtually all except for short sequences (such as individual cosmids) in GenBank were positioned onto the whole-genome BAC map (see Methods). Additional information, such as BAC end sequence alignment and clone sequence overlap, was used to augment the *in silico* digest placement (if needed) and, in some cases, multiple sequences were positioned as part of larger assemblies of overlapping sequences (NT segments, NCBI). From these analyses, we determined that as many as 11.5% of the sequences in GenBank had incorrect clone names referenced in their GenBank records, probably owing mostly to data tracking and clone retrieval errors at the genome centres. A consequence of these naming errors was that many contigs contained clones associated with multiple markers determined by ePCR that mapped collectively to a single region of the genome, but were inconsistent with the remaining clone-to-marker associations in the contig. This was a direct result of incorrect clone names being associated with sequences and hence, incorrect assignment of the markers to those clones in the FPC database. Mapping of sequences to the clone map corrected the naming errors and resolved seemingly out of place ePCR markers once the sequence in which they were detected was properly situated within the ordered contigs. We have found that the correct clone can be retrieved 95% of the time using the whole-genome BAC map, as judged by comparing the fingerprint obtained for the retrieved clone with that in the database. Some clones could not be retrieved owing to growth failures, and others represent data-tracking errors within the fingerprint set. The high level of redundancy of the whole-genome BAC map allows a substitute clone to be readily selected to replace the 5% of clones that are not recovered on the first attempt.

Once the sequences were aligned to the whole-genome BAC fingerprint map, we used these data as a foundation for determining a nonredundant sequence path across the genome. The map order and placement of the sequences with respect to the whole-genome BAC map were considered in the sequence assembly to minimize errors due to potential false assignment of overlaps between related but not identical sequences. The BAC map placements were used as a localization guide only and did not completely constrain the sequence assembly, to avoid any propagation of errors and imprecision of clone placement. The analysis of markers identified within the genome sequence enables a detailed comparison of the whole-genome BAC map with other established landmark-content, radiation hybrid and genetic maps³. There was overall agreement between the sequence assembly that overlays the whole-genome BAC map and other existing maps, with local exceptions. In most instances, these local disagreements indicated the need simply to reverse the current orientation of the underlying BAC contig, and this has been done in the present version of the map.

Conclusions

The whole-genome BAC map allowed the integration of a range of data, including FISH cytogenetic clone localizations, landmark data obtained by PCR and hybridization screening, clones from other libraries with associated map data, and working draft and finished clone sequence and associated ePCR landmarks. New data will continue to be incorporated into this growing database. The entire FPC database of the human genome BAC fingerprint map can be obtained from <http://genome.wustl.edu/gsc/human/Mapping/index.shtml>. A searchable AceDb³⁹ version of the whole-genome BAC map is also accessible at <http://genome.wustl.edu/gsc/Search/db.shtml>, and an overview of the map is available as Supplementary Information.

This clone-based map has been vital for the accurate assembly of the human genome sequence³. The BAC clones comprising the clone-based map also provide an integrated resource for analysis of chromosome structure, comparative genome hybridization⁴⁰ and functional genetics, including gene inactivation⁴¹. Together, the human genome clone map and the anchored sequence map provide synergistic resources for future analysis of the human genome. □

Methods

Regional approach to large-scale physical map construction

The general approach involved screening genomic BAC and PAC libraries by PCR or by probe hybridization using overgo probes⁴² to identify clones corresponding to specific STS markers. Overgo probes are made by filling in the single-stranded overhangs of two overlapping oligonucleotides using radiolabelled nucleotides and Klenow polymerase. Typically, we used two 24-mers overlapping by 8 bp to generate a radiolabelled double-stranded 40-mer. Overgo probes were arranged in three-dimensional arrays with six probes on each axis (giving 216 probes each). A five-directional pooling strategy allowed resolution of 80–90% of all markers with only 30 hybridizations. More than 25,000 human and mouse markers have been associated with BACs using this probe type at the WUGSC (J. McPherson). Once identified, fingerprints were generated from marker-positive clones using *Hind*III restriction enzyme digests with fragment separation on 1% agarose gels⁴³, analysed using Image (<http://www.sanger.ac.uk/Software/Image/>)^{20,21} and the fingerprints examined manually within FPC to build contigs and to select clones for sequencing that span contigs. Manual editing of the automated band calls was required because of inconsistencies in band identification.

Fluorescence *in situ* hybridization

Probes were generated from aliquots of the BAC DNA used to generate the *Hind*III fingerprints using the Prime-it Fluor labelling kit (Stratagene), which incorporates fluor-12-dUTP into the probe fragments by random priming. Probes were hybridized to chromosome spreads on slides with competitor DNA present. Slides were processed essentially according to standard methods⁴⁴. Data were collected and analysed using a Zeiss Axiophot microscope equipped with the Genus camera setup and software (Applied Imaging Corporation).

Integration of sequenced clones using synthetic fingerprints from *in silico* digests

BAC-sized clones were simulated from finished contiguous sequenced regions of DNA.

The sequences were cut into 175-kb fragments each with 40% overlap with the previous segments. Bands less than 600 bp were then removed from consideration to be consistent with the fingerprint data. Fingerprint data were converted from mobilities to sizes and clones from the fingerprinting effort could then be directly compared to sequenced clones from any library or group (when comparing size data, the FPC tolerance variable was set to 10). For clones that were not finished, each contig of the sequence was digested and all end fragments were removed. The remaining fragments were summed to create an *in silico* digest for unfinished clones.

Received 28 November; accepted 27 December 2000.

- Olson, M. V. *et al.* Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl Acad. Sci. USA* **83**, 7826–7830 (1986).
- Coulson, A., Sulston, J., Brenner, S. & Karn, J. Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **83**, 7821–7825 (1986).
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Wilson, R. K. & Mardis, E. R. in *Analyzing DNA* (eds Birren, B., Green, E. D., Klapsholz, S., Myers, R. M. & Roskams, J.) 398–454 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1997).
- Doggett, N. A. *et al.* An integrated physical map of human chromosome 16. *Nature* **377**, 335–365 (1995).
- Ashworth, L. K. *et al.* An integrated metric physical map of human chromosome 19. *Nature Genet.* **11**, 422–427 (1995).
- Crollius, H. R. *et al.* An integrated YAC map of the human X chromosome. *Genome Res.* **6**, 943–955 (1996).
- Bouffard, G. G. *et al.* A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. *Genome Res.* **7**, 673–692 (1997).
- Whitehead Institute for Biomedical Research. YAC STS-content map of the human genome (cited October 2000) (http://carbon.wi.mit.edu:8000/cgi-bin/contig/phys_map) (1997).
- Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
- Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
- Ioannou, P. A. *et al.* A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Genet.* **6**, 84–89 (1994).
- Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA* **89**, 8794–8797 (1992).
- Marra, M. *et al.* A map for sequence analysis of the *Arabidopsis thaliana* genome. *Nature Genet.* **22**, 265–270 (1999).
- Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
- Hoskins, R. A. *et al.* A BAC-based physical map of the major autosomes of *Drosophila melanogaster* [published erratum appears in *Science* **288**, 1751 (2000)]. *Science* **287**, 2271–2274 (2000).
- Pennisi, E. Genomics. Mouse sequencers take up the shotgun. *Science* **287**, 1179–1181 (2000).
- Bouck, J. B., Metzker, M. L. & Gibbs, R. A. Shotgun sample sequence comparisons between mouse and human genomes. *Nature Genet.* **25**, 31–33 (2000).
- Osoegawa, K. *et al.* A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* (in the press).
- Sulston, J. *et al.* Software for genome mapping by fingerprinting techniques. *Comput. Appl. Biosci.* **4**, 125–132 (1988).
- Sulston, J., Mallett, F., Durbin, R. & Horsnell, T. Image analysis of restriction enzyme fingerprint autoradiograms. *Comput. Appl. Biosci.* **5**, 101–106 (1989).
- Soderlund, C., Humphray, S., Dunham, A. & French, L. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.* **10**, 1772–1787 (2000).
- Soderlund, C., Longden, I. & Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13**, 523–535 (1997).
- Schuler, G. D. *et al.* A gene map of the human genome. *Science* **274**, 540–546 (1996).
- Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
- The International Radiation Hybrid Mapping Consortium. A new gene map of the human genome: GeneMap'99. (cited October 2000) (<http://www.ncbi.nlm.nih.gov/genemap>) (1999).
- Zhao, S. *et al.* Human BAC ends quality assessment and sequence analyses. *Genomics* **63**, 321–332 (2000).
- The BAC Resource Consortium. Integration of autogenetic landmarks into the draft sequence of the human genome. *Nature* **409**, 953–958 (2001).
- Schuler, G. D. Electronic PCR: bridging the gap between genome mapping and genome sequencing. *Trends Biotechnol.* **16**, 456–459 (1998).
- Montgomery, K. *et al.* A high-resolution map of human chromosome 12. *Nature* **409**, 945–946 (2001).
- Bruls, T. *et al.* A physical map of human chromosome 14. *Nature* **409**, 947–948 (2001).
- Tilford, C. A. *et al.* A physical map of the human Y chromosome. *Nature* **409**, 943–945 (2001).
- Bentley, D. R. *et al.* The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409**, 942–943 (2001).
- Reithman, H. C. *et al.* Integration of telomere sequences with the draft human genome sequences. *Nature* **409**, 948–951 (2001).
- Asakawa, S. *et al.* Human BAC library: construction and rapid screening. *Gene* **191**, 69–79 (1997).
- The chromosome 21 mapping and sequencing consortium. The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
- Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
- Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **28**, 15–18 (2000).
- Eeckman, F. H. & Durbin, R. ACeDB and macace. *Methods Cell Biol.* **48**, 583–605 (1995).
- Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet.* **20**, 207–211 (1998).
- Capecchi, M. R. Choose your target. *Nature Genet.* **26**, 159–161 (2000).
- Ross, M. T., LaBrie, S., McPherson, J. & Stanton, V. P. in *Current Protocols in Human Genetics* (eds Dracopoli, N. C. *et al.*) 5.6.1–5.6.5 (Wiley, New York, 1999).

43. Marra, M. A. *et al.* High throughput fingerprint analysis of large-insert clones. *Genome Res.* 7, 1072–1084 (1997).
44. Lichter, P., Boyle, A. L., Cremer, T. & Ward, D. C. Analysis of genes and chromosomes by nonisotopic in situ hybridization. *Genet. Anal. Tech. Appl.* 8, 24–35 (1991).
45. Morton, N. E. Parameters of the human genome. *Proc. Natl Acad. Sci. USA* 88, 7474–7476 (1991).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank everyone who has contributed to mapping the human genome by providing data to GenBank and other publicly accessible web sites. As much as possible, these data have been incorporated into the map presented here. The fingerprinting project was funded as

The International Human Genome Mapping Consortium*

Washington University School of Medicine, Genome Sequencing Center: John D. McPherson¹, Marco Marra^{1*}, LaDeana Hillier¹, Robert H. Waterston¹, Asif Chinwalla¹, John Wallis¹, Mandeep Sekhon¹, Kristine Wylie¹, Elaine R. Mardis¹, Richard K. Wilson¹, Robert Fulton¹, Tamara A. Kucaba¹, Caryn Wagner-McPherson¹ & William B. Barbazuk¹

Wellcome Trust Genome Campus: Simon G. Gregory², Sean J. Humphray², Lisa French², Richard S. Evans², Graeme Bethel², Adam Whittaker², Jane L. Holden², Owen T. McCann², Andrew Dunham², Carol Soderlund^{2*}, Carol E. Scott² & David R. Bentley²

National Center for Biotechnology Information: Gregory Schuler³, Hsiu-Chuan Chen³ & Wonhee Jang³

National Human Genome Research Institute: Eric D. Green⁴, Jacquelyn R. Idol⁴ & Valerie V. Braden Maduro⁴

Albert Einstein College of Medicine: Kate T. Montgomery⁵, Eunice Lee⁵, Ashley Miller⁵, Suzanne Emerling⁵ & Raju Kucherlapati⁵

Baylor College of Medicine, Human Genome Sequencing Center: Richard Gibbs⁶, Steve Scherer⁶, J. Harley Gorrell⁶, Erica Sodergren⁶, Kerstin Clerc-Blankenburg⁶, Paul Tabor⁶, Susan Naylor⁷ & Dawn Garcia⁷

Roswell Park Cancer Institute: Pieter J. de Jong^{8*}, Joseph J. Catanese^{8*}, Norma Nowak⁸ & Kazutoyo Osoegawa^{8*}

Multimegabase Sequencing Center: Shizhen Qin⁹, Lee Rowen⁹, Anuradha Madan⁹, Monica Dors⁹ & Leroy Hood⁹

Fred Hutchinson Cancer Research Institute: Barbara Trask¹⁰, Cynthia Friedman¹⁰ & Hillary Massa¹⁰

The Children's Hospital of Philadelphia: Vivian G. Cheung¹¹, Ilan R. Kirsch¹², Thomas Reid¹² & Raluca Yonescu¹²

Genoscope: Jean Weissenbach¹³, Thomas Bruls¹³ & Roland Heilig¹³

US DOE Joint Genome Institute: Elbert Branscomb¹⁴, Anne Olsen¹⁴, Norman Doggett¹⁴, Jan-Fang Cheng¹⁴ & Trevor Hawkins¹⁴

Stanford Human Genome Center and Department of Genetics: Richard M. Myers¹⁵, Jin Shang¹⁵, Lucia Ramirez¹⁵, Jeremy Schmutz¹⁵, Olivia Velasquez¹⁵, Kami Dixon¹⁵, Nancy E. Stone¹⁵ & David R. Cox¹⁵

University of California, Santa Cruz: David Haussler^{16,17}, W. James Kent¹⁸, Terrence Furey¹⁷, Sanja Rogic¹⁷ & Scot Kennedy¹⁹

part of the Human Genome Project sequencing initiative of the NHGRI. The Keio group was supported in part by the Fund for Human Genome Sequencing Project from the JST and the Fund for "Research for the Future" Program from the Japan Society for the Promotion of Science. The RIKEN GSC group is supported by the Special Fund for Human Genome Sequencing from Science and Technology Agency, Japan and a Grant-in-Aid Scientific Research on Priority Area, "Genome Science" from Monbusho, Japan. Multiple US groups were funded by NIH/NCI and DOE. The MPIMG acknowledge grants from the Max-Planck-Society and the Federal German Ministry of Education, Research and Technology (BMBF) through Projektraeger DLR, in the framework of the German Human Genome Project. Data management throughout the project was facilitated by using a suitably modified AceDb database (R. Durbin and J. Thierry-Mieg). For a complete list of all authors, see Supplementary Information.

Correspondence and requests for materials should be addressed to J.D.M. (e-mail: jmcphe@watson.wustl.edu).

British Columbia Cancer Research Centre: Steven Jones²⁰

Department of Genome Analysis, Institute of Molecular Biotechnology: André Rosenthal²¹, Gaiping Wen²¹, Markus Schilhabel²¹, Gernot Gloeckner²¹, Gerald Nyakatura^{21*}, Reiner Siebert²² & Brigitte Schlegelberger²²

Departments of Human Genetics and Pediatrics, University of California: Julie Korenberg²³ & Xiao-Ning Chen²³

RIKEN Genomic Sciences Center: Asao Fujiyama²⁴, Masahira Hattori²⁴, Atsushi Toyoda²⁴, Tetsushi Yada²⁴, Hong-Seok Park²⁴ & Yoshiyuki Sakaki²⁴

Department of Molecular Biology, Keio University School of Medicine: Nobuyoshi Shimizu²⁵, Shuichi Asakawa²⁵, Kazuhiko Kawasaki²⁵, Takashi Sasaki²⁵, Ai Shintani²⁵, Atsushi Shimizu²⁵, Kazunori Shibuya²⁵, Jun Kudoh²⁵ & Shinsei Minoshima²⁵

Max-Planck-Institute for Molecular Genetics: Juliane Ramser²⁶, Peter Seranski^{26,27}, Celine Hoff^{26,27}, Annemarie Poustka^{26,27}, Richard Reinhardt²⁶ & Hans Lehrach²⁶

1, Washington University School of Medicine, Genome Sequencing Center, Department of Genetics, 4444 Forest Park Boulevard, St. Louis, Missouri 63108, USA; 2, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK; 3, National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA; 4, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA; 5, Department of Molecular Genetics, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA; 6, Baylor College of Medicine, Human Genome Sequencing Center, Houston, Texas, USA; 7, University of Texas, San Antonio, Texas, USA; 8, Roswell Park Cancer Institute, Buffalo, New York 14263, USA; 9, Multimegabase Sequencing Center, Institute for Systems Biology, Seattle, Washington 98105, USA; 10, Division of Human Biology, Fred Hutchinson Cancer Research Institute, Seattle, Washington 98109, USA; 11, Department of Pediatrics, The Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; 12, Genetics Department, Medicine Branch, National Cancer Institute, Washington DC, USA; 13, Genoscope, Centre National de Séquencage, 2 Rue Gaston Crémieux, CP 5706, 91057 Evry, France; 14, US DOE Joint Genome Institute, Walnut Creek, California, USA; 15, Stanford Human Genome Center and Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA; 16, Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, California 95064, USA; 17, Department of Computer Science, University of California, Santa Cruz, Santa Cruz, California 95064, USA; 18, Department of Biology, University of California, Santa Cruz, Santa Cruz, California 95064, USA; 19, Department of Mathematics, University of California, Santa Cruz, Santa Cruz, California 95064, USA; 20, British Columbia Cancer Research Centre, 600 West 10th Avenue, Room 3427, Vancouver, British Columbia V5Z 4E6, Canada; 21, Dept. of Genome Analysis, Institute of Molecular Biotechnology, Beutenbergstrasse 11, D-07745 JENA, Germany; 22, Institute of Human Genetics, University of Kiel, Germany; 23, Departments of Human Genetics and Pediatrics, University of California, Los Angeles, California, USA;

24, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan; 25, Department of Molecular Biology, Keio University School of Medicine, 35 Shinanomachi Shinjuku-ku, Tokyo 160-8582, Japan; 26, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, D-14195, Berlin, Germany; 27, Abteilung Molekulare Genomanalyse, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, 69120, Heidelberg, Germany.

* Present addresses: British Columbia Cancer Research Centre, 600 West 10th Avenue, Room 3427, Vancouver, British Columbia V5Z 4E6, Canada (M.M.); Clemson University Genome Institute, 100 Jordan Hall, Clemson University, Clemson, South Carolina 29634-5727, USA (C.S.); Children's Hospital Oakland Research Institute, BACPAC Resources, Oakland, California 94609, USA and Pfizer Global Research & Development, Alameda Laboratories, Alameda, California 94502 USA (P.J.d.J., J.J.C., K.O.); MWG-Biotech AG, Ebersberg, Germany (G.N.).

The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X

D. R. Bentley*, P. Deloukas*, A. Dunham*, L. French*, S. G. Gregory*, S. J. Humphray*, A. J. Mungall*, M. T. Ross*, N. P. Carter*, I. Dunham*, C. E. Scott*, K. J. Ashcroft*, A. L. Atkinson*, K. Aubin*, D. M. Beare*, G. Bethel*, N. Brady*, J. C. Brook*, D. C. Burford*, W. D. Burrill*, C. Burrows*, A. P. Butler*, C. Carder*, J. J. Catanese*, C. M. Clee*, S. M. Clegg*, V. Cobley*, A. J. Coffey*, C. G. Cole*, J. E. Collins*, J. S. Conquer*, R. A. Cooper*, K. M. Culley*, E. Dawson*, F. L. Dearden*, R. M. Durbin*, P. J. de Jong*, P. D. Dhami*, M. E. Earthrowl*, C. A. Edwards*, R. S. Evans*, C. J. Gillson*, J. Ghorl*, L. Green*, R. Gwilliam*, K. S. Halls*, S. Hammond*, G. L. Harper*, R. W. Heathcott*, J. L. Holden*, E. Holloway*, B. L. Hopkins*, P. J. Howard*, G. R. Howell*, E. J. Huckle*, J. Hughes*, P. J. Hunt*, S. E. Hunt*, M. Izmajlowicz*, C. A. Jones*, S. S. Joseph*, G. Laird*, C. F. Langford*, M. H. Lehtvaslaiho*, M. A. Leversha*, O. T. McCann*, L. M. McDonald*, J. McDowall*, G. L. Maslen*, D. Mistry*, N. K. Moschonas*, V. Neocleous*, D. M. Pearson*, K. J. Phillips*, K. M. Porter*, S. R. Prathalingam*, Y. H. Ramsey*, S. A. Ranby*, C. M. Rice*, J. Rogers*, L. J. Rogers*, T. Sarafidou*, D. J. Scott*, G. J. Sharp*, C. J. Shaw-Smith*, L. J. Smink*, C. Soderlund*, E. C. Sothoran*, H. E. Steingruber*, J. E. Sulston*, A. Taylor*, R. G. Taylor*, A. A. Thorpe*, E. Tinsley*, G. L. Warry*, A. Whittaker*, P. Whittaker*, S. H. Williams*, T. E. Wilmer*, R. Wooster* & C. L. Wright*

* The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

‡ Department of Biology, University of Crete and Institute of Molecular Biology and Biotechnology, PO Box 2208, 71409 Heraklion, Crete, Greece

§ Neurogenetic Laboratory, The Cyprus Institute of Neurology and Genetics, 6, International Airport Avenue, PO Box 23462, 1683 Nicosia, Cyprus

† Children's Hospital-BACPAC Resources, 747 52nd Street, Oakland, California 94609, USA

We constructed maps for eight chromosomes (1, 6, 9, 10, 13, 20, X and (previously) 22), representing one-third of the genome, by building landmark maps, isolating bacterial clones and assembling contigs. By this approach, we could establish the long-range organization of the maps early in the project, and all contig extension, gap closure and problem-solving was simplified by containment within local regions. The maps currently represent more than 94% of the euchromatic (gene-containing) regions of these chromosomes in 176 contigs, and contain 96% of the chromosome-specific markers in the human gene map. By measuring the remaining gaps, we can assess chromosome length and coverage in sequenced clones.

The task of sequencing the 3,200 megabase (Mb) human genome

can be subdivided into individual chromosome projects ranging in size from 263 Mb (chromosome 1)¹ to about 35 Mb (chromosomes 21q and 22q)^{2,3}. Our strategy, in common with other groups^{4–6}, was to map selected chromosomes individually, and then to combine the results with those of whole-genome mapping studies into a single map of the human genome⁷. Chromosome maps were constructed as follows (see Supplementary Information). First, we constructed a landmark map for each chromosome. Second, we identified bacterial clones (bacterial- or P1-derived artificial chromosomes (BACs or PACs)) from genomic libraries using the chromosome-specific landmarks, and assembled them into contigs on the basis of shared restriction enzyme fingerprints and landmark content. Third, contigs were extended and joined by chromosome walking. Walking was carried out using BAC end sequences generated in house from clones selected at the ends of contigs, or from the publicly available resources; joins were also made by identification of overlaps using genomic sequence data.

Clones that were representative of each contig were selected regularly for inclusion in the 'tiling path' (a set of minimally overlapping clones) for genomic sequencing⁸. Clones from the tiling path of each chromosome were deposited in the 'HumanMap' database at the Genome Sequencing Centre⁷ (<http://genome.wustl.edu/gsc/human/mapping/>), for integration with the data obtained by whole-genome fingerprinting. New clones were identified from this integrated dataset to assist the extension and closure of the chromosome maps (Table 1). A detailed description of tiling paths and the underlying clone contigs is available as Supplementary Information and will continue to be updated at www.sanger.ac.uk; all clones are publicly available.

A key question is the extent of coverage of each chromosome in the map. We analysed the coverage of the euchromatic regions on the basis of the assumptions given below and the approximate estimates that were available for chromosome length. Heterochromatic regions, which are estimated to comprise 3–15% of each chromosome (Table 1), are absent from the contigs analysed here. On the basis of the fingerprint bands in the maps (converted to Mb as described in Supplementary Information), 176 contigs represent 927 Mb, or 94% of the estimated total of 981 Mb euchromatin. Half of this (485 Mb) is in fifteen contigs of 22–62 Mb, that illustrates the extent of continuity obtained. We also analysed the representation of human gene markers in the map. The clone map contained 96% of the unique markers that were previously mapped to these chromosomes in GeneMap99 (ref. 9), and 90% were present in the genome sequence. As expected, this figure is higher for individual chromosomes 20 (99%) and 22q (98%), which are nearly or completely finished (Table 1).

Independent corroboration of the coverage of each chromosome required identification of the boundaries between euchromatic sequence and the centromeric, telomeric, and other heterochromatic repeat sequences, as well as measurement of the remaining gaps in the map. We have measured 52 of the 57 remaining gaps in

Table 1 Status of chromosome maps

	Estimated length		Physical map			Gene map		
	Total	Euchromatin	Contigs	Euchromatin covered		Markers analysed	Markers in contigs	Markers in sequence
Chromosome	(Mb)	(Mb)	(n)	(Mb)	(%)	(n)	(%)	(%)
1	263	238	78	226	95	4,769	94	87
6	183	178	13	170	96	2,942	96	92
9	145	123	21	113	92	1,771	98	90
10	144	139	14	124	89	2,022	99	91
13	114	98	7	102	105	1,028	98	90
20	72	67	6	61	91	1,159	99	99
22(q12–tel)	25.5	25.5	8	24.8	97	777	98	98
X(p21–q27)	115	112	29	106	94	925	93	88
Total	1,062	981	176	927	94	15,393	96	90

See methods in Supplementary Information for details.

the maps of chromosomes 6, 9, 10, 13 and 20 by fluorescent *in situ* hybridization (FISH), in addition to the gaps previously measured on chromosome 22. Clones immediately flanking each gap were hybridized to extended DNA fibres, interphase nuclei or metaphase chromosome spreads.

An example of this analysis is given for chromosome 10 (Fig. 1). Fourteen contigs account for 124.4 Mb of DNA. They contain pericentromeric satellite sequences, and also a subtelomeric sequence on the short arm which is <0.1 Mb from the telomere; the most distal clone on the long arm is <0.15 Mb from the telomere¹⁰. From FISH analysis, the total extent of euchromatic gaps is ≤ 4.2 Mb. From these measurements, the estimated coverage of euchromatin in the map can be revised from 89% (based on Morton's previous estimate; Table 1) to 96.7% (124.4 of 128.6 Mb). A 9.75-Mb restriction map that spans the centromere has been constructed for chromosome 10 (ref. 11). By anchoring this to the clone maps on both chromosome arms, we estimate that the size of the gap across the centromeric region is around 4.5 Mb. This analysis provides a new estimate of 133 Mb for the total length of the chromosome (Fig. 1), compared with the previous value of 144 Mb. In a similar analysis of chromosome 13, seven contigs span 102 Mb and the remaining gaps in the euchromatic region total a maximum of 5 Mb. This leads to a new, increased estimate of the size of the euchromatic region of 107 Mb, and the current coverage in the map is 95%.

Our strategy may form the basis for finishing the map and sequence of every human chromosome. Accurate measurement of all gap sizes assists our continued efforts to bridge the remaining gaps in clones, using other cloning systems (for example yeast artificial chromosomes (YACs), plasmids and bacteriophage

lambda), and characterization of the centromeric satellites and the remaining heterochromatic regions, including the short arms of the acrocentric chromosomes (13, 14, 15, 21 and 22), will enable us to determine the true physical extent of DNA in the human genome. □

Received 28 November; accepted 21 December 2000.

1. Morton, N. E. Parameters of the human genome. *Proc. Natl Acad. Sci. USA* **88**, 7474–7476 (1991).
2. The chromosome 21 mapping and sequencing consortium. The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
3. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
4. Tilford, C. A. *et al.* A physical map of the human Y chromosome. *Nature* **409**, 943–945 (2001).
5. Montgomery, K. T. A high-resolution map of human chromosome 12. *Nature* **409**, 945–946 (2001).
6. Bruls, T. *et al.* A physical map of human chromosome 14. *Nature* **409**, 947–948 (2001).
7. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
8. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
9. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
10. Riethman, H. C. *et al.* Integration of telomere sequences with the draft human genome sequence. *Nature* **409**, 948–951 (2001).
11. Jackson, M. S., See, C. G., Mulligan, L. M. & Lauffart, B. F. A 9.75-Mb map across the centromere of human chromosome 10. *Genomics* **33**, 258–270 (1996).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank M. Sekhon, A. Chinwalla, J. McPherson and staff of the Genome Sequencing Centre, St. Louis for their assistance; the many collaborators who have contributed reagents and information to assist map construction on individual chromosomes; M. Jackson for helpful discussion; T. Cox, R. Pettett and the web team; and the Wellcome Trust for support.

Correspondence and requests for material should be addressed to D.R.B. (e-mail: drb@sanger.ac.uk).

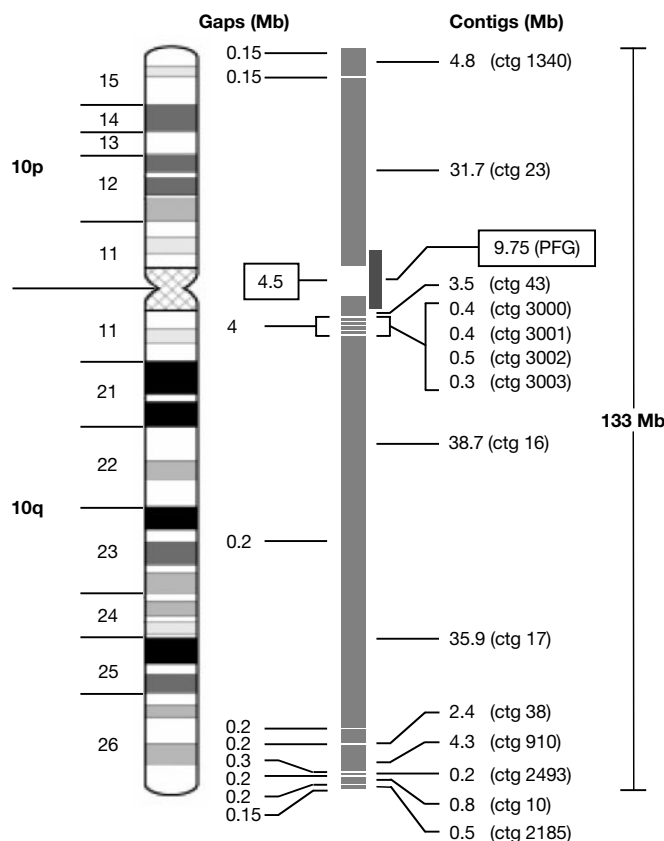


Figure 1 Physical map of chromosome 10. Clone contigs cover 124.4 Mb, euchromatic gaps cover 4.2 Mb, and the gap across centromeric satellites is 4.5 Mb¹¹. The separation between contigs 43 and 16 was determined as ~4 Mb on the basis of FISH of metaphase chromosome spreads, from which the sum of contigs 3000–3003 (1.6 Mb) was subtracted for a net gap coverage of 2.4 Mb in this region of 10q11.

A physical map of the human Y chromosome

Charles A. Tilford*, Tomoko Kuroda-Kawaguchi*, Helen Skaletsky*, Steve Rozen*, Laura G. Brown*, Michael Rosenberg*, John D. McPherson†, Kristine Wylie†, Mandeep Sekhon†, Tamara A. Kucaba†, Robert H. Waterston† & David C. Page*

* Howard Hughes Medical Institute, Whitehead Institute, and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA

† Genome Sequencing Center, Department of Genetics, Washington University School of Medicine, 4444 Forest Park Boulevard, St. Louis, Missouri 63108, USA

The non-recombining region of the human Y chromosome (NRY), which comprises 95% of the chromosome, does not undergo sexual recombination and is present only in males. An understanding of its biological functions has begun to emerge from DNA studies of individuals with partial Y chromosomes, coupled with molecular characterization of genes implicated in gonadal sex reversal, Turner syndrome, graft rejection and spermatogenic failure^{1,2}. But mapping strategies applied successfully elsewhere in the genome have faltered in the NRY, where there is no meiotic recombination map and intrachromosomal repetitive sequences are abundant³. Here we report a high-resolution physical map of the euchromatic, centromeric and heterochromatic regions of the NRY and its construction by unusual methods, including genomic clone subtraction⁴ and dissection of sequence family variants⁵. Of the map's 758 DNA markers, 136 have multiple locations in the NRY, reflecting its unusually repetitive sequence composition. The markers anchor 1,038 bacterial artificial chromosome clones, 199 of which form a tiling path for sequencing.

A low-resolution physical map of the Y chromosome was previously assembled by testing naturally occurring deletions and yeast artificial chromosome (YAC) clones for the presence or absence of 182 Y-chromosomal sequence-tagged sites (STSs)^{3,6}. These STS markers were generated from Y DNA sequences selected at random, which promoted representative sampling of the entire chromosome⁶. Nonetheless, most randomly selected sequences proved unusable as map landmarks because they corresponded either to interspersed repetitive elements found throughout the genome or to male-specific repetitive sequences dispersed to many locations in the NRY.

To construct a high-resolution map, we generated additional STSs in a directed manner. To enrich for the single-copy sequences most useful as map landmarks, we systematically applied genomic clone subtraction, whereby a 'tracer' clone's DNA is depleted of sequences shared with a set of 'driver' clones⁴. We identified a tiling path of 57 YACs that collectively spanned the euchromatic NRY³, and then carried out, in parallel, 57 subtractions, each employing one YAC as tracer and the remaining YACs (minus those overlapping the tracer YAC) as drivers. We sequenced a random sample of products from each of the 57 subtractions, identifying 308 additional STSs that proved useful in map assembly.

We used radiation hybrid mapping to integrate and order the random and subtraction-derived STSs. Large-fragment radiation hybrid panels offering long-range connectivity have been used to assemble human genome maps at 500–1,000-kilobase (kb) resolution^{7–10}. To obtain greater resolution in the NRY, where the number of STSs appeared sufficient to cover the euchromatic region at an

average spacing of 50 kb, we used a small-fragment radiation hybrid panel that had been used to build detailed maps of limited autosomal segments¹¹. We tested this panel for all random and subtraction-derived NRY STSs. Nascent radiation hybrid linkage groups were ordered and oriented with respect to the centromere by positioning selected STSs on the existing map of natural deletions⁶. Additional STSs generated in our project's later phases were also tested. Ultimately, 513 STSs were positioned, at a resolution of around 50 kb, on a radiation hybrid map encompassing nearly the entire euchromatic NRY.

To prepare for sequencing the NRY, we used the radiation hybrid map as a scaffold for assembling contigs of bacterial artificial chromosome (BAC) clones. We screened a BAC library of human male genomic DNA with hybridization probes derived from NRY STSs. Through subsequent polymerase chain reaction tests of STS content, we assembled 1,038 BACs into contigs that, except for four small gaps, represented the whole NRY (see Fig. 1 in Supplementary Information). Many portions of this BAC map could be assembled only after the sequence of selected BACs had been determined and compared. Also, many NRY genes and extragenic sequences are known to have closely related counterparts on the X chromosome¹². In many cases, it was initially unclear whether BACs identified using X-homologous NRY STSs, especially those from a 4-Mb region of 99% X–Y identity^{13,14}, derived from the NRY or from the X chromosome. We resolved these ambiguities by resequencing STSs from the BACs in question and comparing them to X- or Y-derived reference sequences. The resulting map of overlapping BACs and ordered STSs (Fig. 1 in Supplementary Information)

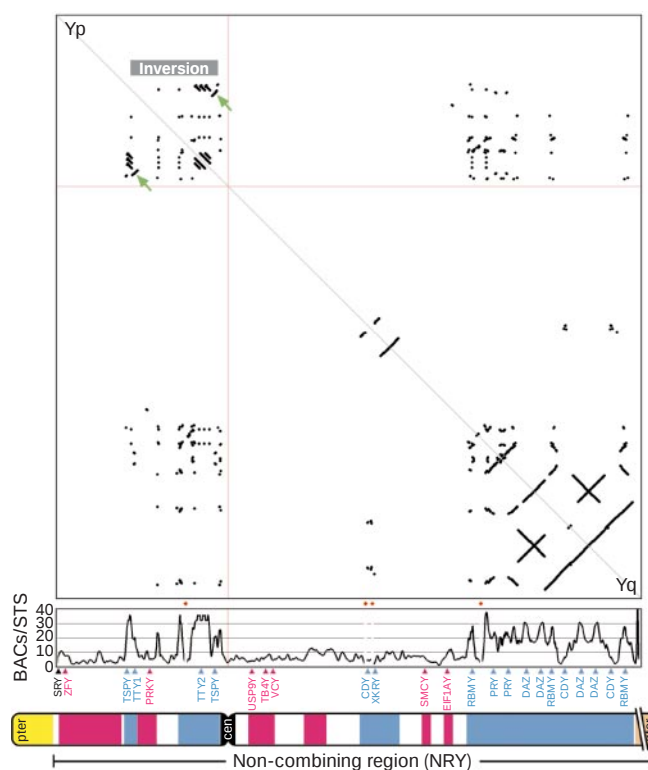


Figure 1 Repetitive structure of euchromatic NRY. Bottom, schematic of the Y chromosome, comprising large NRY flanked by pseudoautosomal regions (yellow). NRY is divided into euchromatic and heterochromatic (tan, shown truncated) portions, roughly 24 and 30 Mb, respectively. pter, short-arm telomere; cen, centromere; qter, long-arm telomere. Within euchromatic NRY, regions rich in NRY-specific amplicons (blue) or sequence similarity to X chromosome (red) are shown. Above chromosome schematic are positions of some NRY genes; most are found in amplicons (blue) or have X-linked homologues (red). Above genes is a plot of the average number of NRY BACs that contain each of the 758 STSs mapped (136 of these STSs at two or more locations) along euchromatic NRY. As expected, STSs in amplicon regions tended to be present in more BACs than STSs in X-homologous or

unshaded regions. (Plotted values are local averages within sliding window of five consecutive STSs; values reflect all NRY BACs containing those STSs, not just BACs assigned to site indicated.) Some amplicon regions were under-represented in the BAC library; four gaps remain (red diamonds; ≤ 100 kb each) in BAC coverage of NRY. Top, STS-based dot plot of euchromatic NRY. Each dot reflects occurrence of a particular STS at two points in map (complete map shown in Fig. 1 in Supplementary Information). Dots fall almost exclusively within amplicon regions. Many repeats of entire groups of STSs are apparent, with lines parallel to light grey diagonal indicating direct repeats and lines perpendicular to light grey diagonal indicating inverted repeats. Green arrows, inverted repeats flanking 3.5-Mb inversion (see text). Pale red lines, centromere.

was extensively cross-checked against the radiation hybrid map and was further reinforced by restriction fingerprinting of all mapped BACs¹⁵.

The greatest challenges were posed by massive, NRY-specific amplified regions (or amplicons), which comprise about one-third of the euchromatic NRY. Of the 758 STSs on which the map is built, 136 are present at two or more locations in the NRY. Although we avoided such repetitive STSs in favour of single-copy STSs wherever possible, substantial portions of the euchromatic NRY contained little or no single-copy sequence. For many such amplicons, BACs derived from different copies could not be distinguished by STS content or restriction fingerprinting¹⁵. In many cases, we distinguished among amplicon copies (and the BACs corresponding to them) by typing 'sequence family variants' (SFVs)⁵. SFVs are subtle differences (for example, single-nucleotide substitutions or dinucleotide repeat length alterations) between closely related but non-allelic sequences. We were analysing BACs from only one male's Y chromosome, so these subtle sequence differences could not represent allelic variants. In general, we identified SFVs only after comparing the DNA sequences of BACs that originated from distinct copies, despite having similar STS content. Thus, mapping and sequencing were inseparable, iterative activities in amplicon-rich regions.

The euchromatic NRY amplicons are diverse in composition, size, copy number and orientation (Fig. 1), with some occurring as tandem repeats, others as inverted repeats, and still others dispersed throughout both arms of the chromosome. The euchromatic amplicons are well populated with testis-specific gene families that may be critical in spermatogenesis (see Fig. 1 in Supplementary Information)^{1,2}.

One pair of amplicons is of particular interest in the context of human variation. Highlighted in Fig. 1 (arrows) are two units, each 300 kb long, that exist in opposite orientations on the short arm. These inverted repeats bound a region of around 3.5 Mb that occurs in one orientation (Fig. 1 in Supplementary Information) in the male from whom the BAC library was constructed, but in the opposite orientation in the existing map of naturally occurring deletions⁶. This may reflect variation among men for a 3.5-Mb inversion^{1,16}, perhaps the result of homologous recombination between the 300-kb inverted repeats flanking the inverted segment. Large Y-chromosome inversions are postulated to have been crucial in the evolution of the human sex chromosomes¹², and this 3.5-Mb inversion may be one of many massive NRY variants that exist in modern populations. □

Received 16 November; accepted 21 December 2000.

- Vogt, P. H. *et al.* Report of the Third International Workshop on Y Chromosome Mapping 1997. *Cytogenet. Cell Genet.* **79**, 1–20 (1997).
- Lahn, B. T. & Page, D. C. Functional coherence of the human Y chromosome. *Science* **278**, 675–680 (1997).
- Foot, S., Vollrath, D., Hilton, A. & Page, D. C. The human Y chromosome: Overlapping DNA clones spanning the euchromatic region. *Science* **258**, 60–66 (1992).
- Reijo, R. *et al.* Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene. *Nature Genet.* **10**, 383–393 (1995).
- Saxena, R. *et al.* Four DAZ genes in two clusters found in the AZFc region of the human Y chromosome. *Genomics* **67**, 256–267 (2000).
- Vollrath, D. *et al.* The human Y chromosome: A 43-interval map based on naturally occurring deletions. *Science* **258**, 52–59 (1992).
- Hudson, T. J. *et al.* An STS-based map of the human genome. *Science* **270**, 1945–1954 (1995).
- Gyapay, G. *et al.* A radiation hybrid map of the human genome. *Hum. Mol. Genet.* **5**, 339–346 (1996).
- Stewart, E. A. *et al.* An STS-based radiation hybrid map of the human genome. *Genome Res.* **7**, 422–433 (1997).
- Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
- Lunetta, K. L., Boehnke, M., Lange, K. & Cox, D. R. Selected locus and multiple panel models for radiation hybrid mapping. *Am. J. Hum. Genet.* **59**, 717–725 (1996).
- Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).
- Mumm, S., Molini, B., Terrell, J., Srivastava, A. & Schlessinger, D. Evolutionary features of the 4-Mb Xq21.3 XY homology region revealed by a map at 60-kb resolution. *Genome Res.* **7**, 307–314 (1997).
- Schwartz, A. *et al.* Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. *Hum. Mol. Genet.* **7**, 1–11 (1998).

- The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
- Jobling, M. A. *et al.* A selective difference between human Y-chromosomal DNA haplotypes. *Curr. Biol.* **8**, 1391–1394 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank C. Nusbaum and T. Hudson for materials and advice on radiation hybrid mapping; J. Crockett, J. Fedele, N. Florence, H. Grover, C. McCabe, N. Mudd, S. Sasso, D. Scheer, R. Seim and P. Shelby for technical contributions; and D. Berry, J. Bradley, Y. Lim, A. Lin, D. Menke, M. Royce-Tolland, J. Saionz and J. Wang for comments on the manuscript. Supported in part by NIH.

Correspondence and requests for materials should be addressed to D.C.P. (e-mail: dcpage@wi.mit.edu).

A high-resolution map of human chromosome 12

Kate T. Montgomery*, Eunice Lee*, Ashley Miller*, Stephanie Lau*, Cecilia Shim*, Jeremy Decker*, Denise Chiu*, Suzanne Emerling*, Mandeep Sekhon†, Rachel Kim*, Jack Lenz*, Jinghua Han*, Ilya Ioshikhes*, Beatrice Renault*, Ivonne Marondel*, Sung-Joo Kim Yoon‡, Kyuyoung Song§, V. V. S. Murty||, Steven Scherer¶, Raluca Yonescu#, Ilan R. Kirsch#, Thomas Ried#, John McPherson†, Richard Gibbs¶ & Raju Kucherlapati*

*Department of Molecular Genetics, Albert Einstein College of Medicine, Bronx, New York 10461, USA

‡Catholic University Medical College Research Institutes of Medical Sciences, Seoul 137-404, Korea

§University of Ulsan College of Medicine, Seoul 138-736, Korea

||Department of Pathology, Columbia University College of Physicians and Surgeons, New York, New York 10032, USA

¶Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA

#Genetics Department, Medicine Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

†Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63108, USA

Our sequence-tagged site-content map of chromosome 12 is now integrated with the whole-genome fingerprinting effort^{1,2}. It provides accurate and nearly complete bacterial clone coverage of chromosome 12. We propose that this integrated mapping protocol serves as a model for constructing physical maps for entire genomes.

Human chromosome 12 has been estimated to be 125 megabases (Mb) in size. Several low-density clone maps, genetic linkage maps and radiation hybrid maps that include chromosome 12 have been described^{3–5}. The map presented here is based on large bacterial clones that include bacterial artificial chromosomes (BACs), P1 (phage) artificial chromosomes (PACs; <http://www.chori.org/bacpac>) and a few cosmids⁶. To construct the map, we divided the chromosome into 1–2 centiMorgan (cM) intervals and assembled all sequence-tagged site (STS) markers known to be in each interval. We used probes derived from the non-polymorphic markers in an interval to screen colony arrays representing six to ten genome equivalents from the RPCI libraries. To extend contigs and link adjacent contigs, we used STSs corresponding to end sequences of BACs at the outer boundaries of each contig to screen members of the contig from which they were derived as well as possible adjacent contigs. This process was reiterated until the region was covered by

overlapping clones⁷. The STS-content map was assembled manually and the data stored in a relational database. A representative portion of the map corresponding to 5 cM of 12p can be seen in Supplementary Information; the entire map is available on our website (<http://sequence.aecom.yu.edu/chr12/>).

Our map contains 3,090 STS markers. Of these, 438 are polymorphic, 1,836 are non-polymorphic and the remaining 816 are based on genes or expressed sequence tags (ESTs). These markers were obtained from gene maps⁸, linkage maps^{3,9,10} or the Whitehead Institute radiation hybrid map (<http://www.genome.wi.mit.edu>), or developed at the Albert Einstein College of Medicine Genome Center. If the size of chromosome 12 is 125 Mb, the STS markers provide an average resolution of 40 kb. There are more than 5,300 large-insert clones on the map, and each marker is present in an average of 8.1 bacterial clones. This depth provides a high level of confidence in the quality of the map. Originally, 1,154 tiling path and seeder clones were identified for sequencing chromosome 12; 1,115 of them are being sequenced. The map indicates that some of the seeder clones are redundant and do not need to be finished. The estimated minimal set of clones covering chromosome 12 is 1,025. On average, adjacent clones overlap by 30%. As the average size of the clones in the RPCI-11 library is 174 kb, we can estimate the physical size of chromosome 12 excluding the gaps (see Supplementary Information) to be 125 Mb.

We assessed the accuracy of the map in two ways. First, we analysed tiling path clones with any available sequence to determine whether they overlapped as predicted by the map. Using electronic polymerase chain reaction (ePCR), BLAST (<http://www.ncbi.nlm.nih.gov/>) or the annotation of finished clones, most overlaps were confirmed. Second, we analysed clones located at 1-Mb intervals along the chromosome using fluorescence *in situ* hybridization (FISH) as part of the Cancer Chromosome Aberration Project (CCAP)^{11,12}. We chose 120 clones, and their order proved to be consistent with the order in the integrated map. The names of the clones, their cytogenetic locations, and details of their distribution are available at <http://www.ncbi.nlm.nih.gov/CCAP/>.

The STS-content map of chromosome 12 has been compared and integrated with the whole-genome map generated by BAC clone fingerprinting. The RPCI-11 BAC library was the primary resource for both maps, so the two are fully compatible. The comparison validated both methods and revealed them to be complementary. When the comparison was initiated, the STS-content map contained 74 contigs anchored to the genetic map. The whole genome map contained 75 contigs anchored to the G4 radiation hybrid map⁴ and nine additional contigs that were assigned to chromosome 12 but unanchored.

The integrated map now consists of 18 contigs that are not linked by PCR or sequence data (see Supplementary Information). The genetic markers on the STS map provide strong evidence for the orientation and order of most of these 18 contigs. This permits the direct comparison of fingerprints of end clones that are adjacent to each other on the map. In two cases we find that the fingerprints overlap and consider these contigs to be manually closed²; in one case we know by sequence that the clones do not overlap and in one the fingerprints are difficult to interpret (see below). The remaining 12 pairs of end clones do not have similar fingerprints and additional clones may be required to achieve contiguity. By these criteria, we consider the chromosome 12 map to have 14 gaps (see Supplementary Information), excluding the centromere, in its bacterial clone coverage.

We previously reported a yeast artificial chromosome (YAC) map of chromosome 12 (ref. 5). When the YAC, BAC and whole-genome fingerprint maps are combined, YACs cover all but five of the gaps, so we expect that these gaps are less than 1 Mb. The gaps not covered by YACs include one in a repetitive region at 43.6 cM (described below), and four between 137.5 and 169.1 cM.

Some of the gaps differ, and may reflect significant genomic

features. Understanding these regions will benefit the mapping and sequencing of this and other genomes. For example, the gap at 13.9 cM is covered by a YAC, but we have been unable to identify a bacterial clone to fill it, despite screening the libraries RPCI-1, 3, 4, 5, 11 and 13 several times with different probes. Furthermore, sequence is currently available for most of the clones in the region, but there are no sequence hits in any of the databases for the ends of the gap. This is the only region of the chromosome that appears to be missing from the available bacterial clone libraries, suggesting that true deficiencies in the bacterial libraries may be rare.

Two gaps appear in the region corresponding to 43.6 cM on 12p. Screening of the 12X BAC library with markers in this region yielded a set of clones whose number far exceeded unique representation of this region in the genome. STS-content analysis of the clones showed that each marker was positive for as many as 30–40 clones and the order of the clones could not be determined logically. These clones had similar restriction patterns and fell into very dense fingerprint contigs that included clones containing markers on chromosomes 2, 3, 11, 12 and 14. The combined fingerprint and PCR results indicate that this region may contain large duplications reminiscent of those described on chromosome 22 (refs 13,14) and it may also represent a region duplicated in other parts of the genome. The fingerprints of the clones flanking one gap in this region have been examined and the data strongly support coverage. Neither fingerprint nor STS data can currently resolve the second gap. This repetitive region may represent one of the most difficult types of region to map and sequence in full, in both the human and other genomes.

This map is one of the most complete human chromosome maps available. The known marker order for each clone allows easy assembly of draft sequence and the map will be invaluable in directing efforts towards completing the sequencing of chromosome 12 as well as future functional genomic studies. □

Received 16 November; accepted 21 December 2000.

- Marra, M. A. *et al.* High throughput fingerprint analysis of large-insert clones. *Genome Res.* **11**, 1072–1084 (1997).
- The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
- Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
- Stewart, E. A. *et al.* An STS-based radiation hybrid map of the human genome. *Genome Res.* **7**, 422–433 (1997).
- Krauter, K. *et al.* A second generation YAC contig map of human chromosome 12. *Nature* **377** (Suppl.), 321–334 (1995).
- Montgomery, K. T. *et al.* Characterization of two chromosome 12 cosmid libraries and development of STSs from cosmids mapped by FISH. *Genomics* **17**, 682–693 (1993).
- Renault, B. *et al.* A sequence-ready physical map of a region of 12q24. 1. *Genomics* **45**, 271–278 (1997).
- Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
- Murray, J. C. *et al.* A comprehensive human linkage map with centimorgan density. *Science* **265**, 2049–2054 (1994).
- Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
- Kirsch, I. R. *et al.* A systematic, high-resolution linkage of the cytogenetic and physical maps of the human genome. *Nature Genet.* **24**, 339–340 (2000).
- Ried, T., Baldini, A., Rand, T. C. & Ward, D. C. Simultaneous visualization of seven different DNA probes by *in situ* hybridization using combinatorial fluorescence and digital imaging microscopy. *Proc. Natl Acad. Sci. USA* **89**, 1388–1392 (1992).
- Edelmann, L. *et al.* A common molecular basis for rearrangement disorders on chromosome 22q11. *Hum. Mol. Genet.* **8**, 1157–1167 (1999).
- Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).

Supplementary information is available from Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

This work is supported by the NIH and the Human Genetics program at Einstein. The Cancer Chromosome Aberrations Project (CCAP) is funded through the office of the director, NCI. We appreciate comments on the manuscript from F. Collins and the assistance of J. Goltz. This work was inspired by F. Ruddle.

Correspondence and requests for materials should be addressed to R.K. (e-mail: kucherla@aecom.yu.edu).

A physical map of human chromosome 14

Thomas Brls*, Gabor Gyapay*, Jean-Louis Petit*,
Franois Artiguenave*, Virginie Vico*, Shizen Qin†,
Aye Mon Tin-Wollam‡, Corinne Da Silva*, Delphine Muselet*,
Delphine Mavel*, Eric Pelletier*, Michael Levy*, Asao Fujiyama§,
Fumihiko Matsuda||, Richard Wilson‡, Lee Rowen†, Leroy Hood†,
Jean Weissenbach*, William Saurin* & Roland Heilig*

* Genoscope and CNRS UMR 8030, 2 Rue Gaston Cremieux, CP 5706,
91057 Evry Cedex, France

† Multimegabase Sequencing Center, The Institute for Systems Biology,
4225 Roosevelt Way, NE Suite 200, Seattle, Washington 98105, USA

‡ Washington University Genome Sequencing Center, Box 8501, 4444 Forest Park
Avenue, St. Louis, Missouri 63108, USA

§ RIKEN Genome Sciences Center, RIKEN, 1-7-22, Tsurumi-ku, Yokohama-city,
Kanagawa 230-0045, Japan

|| Centre National de Genotypage, 2 Rue Gaston Cremieux, CP 5721, 91057 Evry
Cedex, France

We report the construction of a tiling path of around 650 clones covering more than 99% of human chromosome 14. Clone overlap information to assemble the map was derived by comparing fully sequenced clones with a database of clone end sequences^{1,2} (sequence tag connector strategy). We selected homogeneously distributed seed points using an auxiliary high-resolution radiation hybrid map comprising 1,895 distinct positions. The high long-range continuity and low redundancy of the tiling path indicates that the sequence tag connector approach compares favourably with alternative mapping strategies.

We have constructed a map of contiguous DNA fragments cloned in *Escherichia coli* and covering more than 99% of human chromosome 14. Unlike the 'map first and sequence second' approach used for previous chromosome sequencing efforts^{3,4}, the construction of this map was mainly based on a sequence tag connector (STC) strategy^{1,2} in which the map progresses in parallel with the sequencing project. In this iterative approach, fully sequenced bacterial artificial chromosomes (BACs) are searched against a database of clone end sequences to identify minimally overlapping clones and select the next BACs to enter the sequencing pipeline. We combined this 'map as you go' strategy with a dense high-resolution radiation hybrid map. This conjunction conferred a high degree of flexibility on the project: it allowed us to increase the number of relatively evenly distributed seed points and to tailor our efforts in specific chromosomal regions when needed. The fingerprint-based mapping strategy reported by McPherson⁵ also successfully faced challenging schedules and integrated alternative sources of map and sequence data; but we feel that the STC-based approach required less manual editing and resulted in a more accurate and complete set of overlapping clones.

The establishment of ordered sets of clones (contigs) relied on a public database of BAC end sequences² (ftp://ftp.tigr.org/pub/data/h_sapiens/bac_end_sequences/) augmented with end sequences generated in house. Auxiliary mapping resources included a high-resolution chromosome 14 radiation hybrid map and a chromosome 14 enriched library of about 10,000 BACs.

Initially, we chose around 50 seed markers, ordered at high odds and distributed as evenly as possible on the chromosome, using radiation hybrid data from the Genebridge4 panel⁶. We used probes corresponding to the seeds to isolate clones by hybridization against high-density filters of the chromosome 14 BAC library. To cope with the increasing sequencing throughput, the

number of seeds was subsequently tripled by choosing additional BACs that could be mapped unambiguously. We assessed the identity and integrity of selected clones by analysing their restriction fingerprints using the FPC software⁷. Walking could then proceed iteratively and bidirectionally from each fully sequenced seed clone by querying the BAC end sequence database. The walking process entailed systematic checks consisting of re-sequencing of candidate clone ends as well as comparison of their fingerprints.

We monitored the progress of chromosome coverage on a radiation hybrid map built with the TNG high-resolution panel (http://www-shgc.stanford.edu/Mapping/rh/RH_poster/). Roughly 2,350 markers were analysed on this panel, including 640 markers derived from the ends of sequenced clones and 616 entries downloaded from the RHdb database (<http://www.ebi.ac.uk/RHdb>). Marker ordering was treated as being analogous to the well studied 'travelling salesman problem' for which powerful heuristic tools are available⁸ (<http://www.caam.rice.edu/keck/concorde.html>). We then estimated the distances between markers adjacent in the resulting orders using standard maximum likelihood techniques⁹. Information from lower resolution maps was used to improve long-range continuity. The TNG map enabled us to order and orient clone contigs, select additional candidate seed points and determine the gap sizes in the late stages of the project.

Contigs established earlier and covering around 15% of the chromosome were also included in the BAC map. Overall, we selected 162 seed clones (~0.30 chromosome 14 equivalents). About 650 clones were assembled in three clone contigs covering more than 99% of chromosome 14 and totalling around 88 megabases (Mb). The largest contig (~85 Mb) physically connects the most proximal genetic marker (D14S261) to the second most telomeric cluster of markers (D14S292), encompassing almost the entire genetic map of the chromosome. We estimate that fewer than five BACs remain to be incorporated to reach full chromosomal coverage.

Publicly available fingerprint data (ref. 5 and <http://genome.wustl.edu/gsc/human/Mapping>) provided a consistency check for our clone map, which was built independently. Fingerprinted contigs were anchored on our map in several ways. First, fully sequenced clones were digested '*in silico*' and the resulting theoretical restriction pattern aligned against the experimental patterns of the fingerprint clone map⁵. Second, anchoring of fingerprinted contigs was based on the map data associated with markers (including end sequences) contained in the digested clones. These map comparisons indicated that the consistency between fingerprint contigs and our clone scaffold were consistent at a coarse level (some contig junctions were predicted). However, the present clone scaffold was notably more accurate at the finer resolution of local clone ordering and clone overlaps.

The overlap between consecutive BACs resulting from the walking process was estimated to average 20 kilobases (kb) per walking step, and the overlap resulting from random gap closure was estimated to average 52 kb per contig junction. The fraction of redundant sequence was estimated to be 22%, approximately half of which was attributable to suboptimal gap closures.

Note that: (1) most of the redundancy in the clone tiling path is a consequence of the high number of seeds that were required to complete the map in a short time frame¹⁰; (2) this amount of redundancy is still significantly smaller than that computed for typical draft chromosomes (<http://genome.ucsc.edu/>); and (3) the fact that fewer than 1% of the selected clones originated from chromosomes other than 14 shows that the strategy is robust with respect to false links of various origins. This contrasts with the fingerprint-based selection of clones, which led to incorrect chromosomal assignment more frequently.

The distance from the telomere was estimated to be about 5 kb on the basis of fragment restriction data involving a yeast artificial

chromosome clone containing the 14q telomere (F.M., unpublished data). The distance to the alphoid centromeric repeats is unknown. However, the current most centromeric BAC already extends 1,200 kb beyond the most proximal marker of previously reported maps^{11–13}. Interestingly, this clone contains two markers from our TNG map that exhibit extremely high retention rates in the hybrid lines, indicating that they may be close to the centromere.

The clone coverage of chromosome 14 that has been achieved using essentially an STC strategy is very satisfactory, and compares favourably with the coverage obtained for the human chromosomes that have been completely sequenced^{3,4}. The two remaining gaps are located in the subtelomeric part of 14q; subtelomeric regions of many chromosomes are under-represented in most genomic libraries and hence often contain cloning gaps. The largest gap, estimated to be around 600 kb, was subsequently divided into two smaller gaps following the identification of clones through library screening with probes mapped within the gap. The second and more distal gap (20 kb) occurs in the immunoglobulin heavy-chain constant gene region, which contains a number of nearly identical genes and pseudogenes.

Considerations of the optimum design of an STC strategy should include theoretical aspects which correlate the effective depth of the clone end library and the number of seed points to the level of sequence redundancy^{10,14}, as well as practical aspects such as the sequencing capacity and costs¹⁵, the time schedule for the project and the resolution of the available mapping data. However, a centralized repository of BAC end sequences is the only prerequisite for the construction of a tiling path based on the STC approach. Other mapping resources used in this project were auxiliary and provided useful information for seed selection and validation of map extension. Such a strategy is therefore generally portable to any large-scale sequencing project and is readily compatible with partitioning of the project. □

Received 20 October; accepted 21 December 2000.

- Venter, J. C., Smith, H. O. & Hood, L. A new strategy for genome sequencing. *Nature* **381**, 364–366 (1996).
- Mahairas, G. G. *et al.* Sequence-tagged connectors: a sequence approach to mapping and scanning the human genome. *Proc. Natl Acad. Sci. USA* **96**, 9739–9744 (1999).
- Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
- The chromosome 21 mapping and sequencing consortium. The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
- The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
- Gyapay, G. *et al.* A radiation hybrid map of the human genome. *Hum. Mol. Genet.* **5**, 339–346 (1996).
- Soderlund, C., Longden, I. & Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13**, 523–535 (1997).
- Agarwala, R., Applegate, D. L., Maglott, D., Schuler, G. D. & Schaffer, A. A fast and scalable radiation hybrid map construction and integration strategy. *Genome Res.* **10**, 350–364 (2000).
- Lange, K., Boehnke, M., Cox, D. R. & Lunetta, K. L. Statistical methods for polyploid radiation hybrid mapping. *Genome Res.* **5**, 136–150 (1995).
- Roach, J. C., Siegel, A. F., van den Engh, G., Trask, B. & Hood, L. Gaps in the Human Genome Project. *Nature* **401**, 843–845 (1999).
- Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
- Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
- Dear, P. H., Bankier, A. T. & Piper, M. B. A high-resolution metric HAPPY map of human chromosome 14. *Genomics* **48**, 232–241 (1998).
- Batzoglou, S., Berger, B., Mesirov, J. & Lander, E. S. Sequencing a genome by walking with clone-end sequences: a mathematical analysis. *Genome Res.* **9**, 1163–1174 (1999).
- Siegel, A. F. *et al.* Analysis of sequence-tagged-connector strategies for DNA sequencing. *Genome Res.* **9**, 297–307 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank D. Cox, P. Dear and D. Cox for unpublished data and helpful discussions.943

Correspondence and requests for materials should be addressed to R.H. (e-mail: heilig@genoscope.cns.fr).

Integration of telomere sequences with the draft human genome sequence

H. C. Riethman*, Z. Xiang*, S. Paul*, E. Morse*, X.-L. Hu*, J. Flint†, H.-C. Chi‡, D. L. Grady‡ & R. K. Moyzis‡

* The Wistar Institute, 3601 Spruce Street, Philadelphia, Pennsylvania 19104, USA

† Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DS, UK

‡ Department of Biological Chemistry, College of Medicine, University of California, Irvine, California 92697, USA

Telomeres are the ends of linear eukaryotic chromosomes. To ensure that no large stretches of uncharacterized DNA remain between the ends of the human working draft sequence and the ends of each chromosome, we would need to connect the sequences of the telomeres to the working draft sequence. But telomeres have an unusual DNA sequence composition and organization that makes them particularly difficult to isolate and analyse. Here we use specialized linear yeast artificial chromosome clones, each carrying a large telomere-terminal fragment of human DNA, to integrate most human telomeres with the working draft sequence. Subtelomeric sequence structure appears to vary widely, mainly as a result of large differences in subtelomeric repeat sequence abundance and organization at individual telomeres. Many subtelomeric regions appear to be gene-rich, matching both known and unknown expressed genes. This indicates that human subtelomeric regions are not simply buffers of nonfunctional 'junk DNA' next to the molecular telomere, but are instead functional parts of the expressed genome.

Telomeres are essential for genome stability and faithful chromosome replication. The chromatin structures associated with telomeric DNA mediate the many biological activities associated with telomeres, including cell-cycle regulation, cellular ageing, movement and localization of chromosomes within the nucleus, and transcriptional regulation of subtelomeric genes^{1,2}. Specialized functions involving telomeric and subtelomeric DNA have evolved in several eukaryotes. For example, frequent subtelomeric gene conversion provides diversity for surface antigens in trypanosomes³, and rapidly evolving subtelomeric gene families confer selective advantages for closely related yeast strains⁴.

Human telomeres end with a stretch of the conserved simple repeat sequence (TTAGGG)*n*⁵. This tract is present at the end of all telomeres and therefore cannot be used to distinguish one telomere from another. To capture single-copy human DNA regions linked to telomeres that are useful for this purpose, we isolated large telomere-terminal fragments of human chromosomes using specialized yeast artificial chromosome (YAC) cloning vehicles called half-YACs⁶. Each half-YAC clone contains a large segment of subtelomeric DNA flanked by the cloning vector sequence at one end and the human telomere repeat sequence, which has been modified to operate as a functional yeast telomere *in vivo*, at the other. Characterization of these clones revealed low-copy subtelomeric repeats adjacent to the (TTAGGG)*n* sequence^{6,7}. Physical mapping experiments on a large group of these half-YAC clones showed that, in most cases, they can stably maintain faithful copies of human telomere-terminal DNA fragments in yeast⁸. By contrast, bacterial artificial chromosome (BAC) libraries used to prepare the human working draft sequence are not expected to contain sequences extending to the telomere, owing to the absence of restriction sites in (TTAGGG)*n*, the effects of length associated with the construction of size-selected DNA recombinant clones, and the genomic instability of these regions⁹.

We used a combination of chromosome-specific single-copy sequences derived from the half-YAC clones and DNA end sequence derived from cosmid subclones of the half-YACs to connect most telomeres to the working draft sequence (Fig. 1). Our results show that the working draft sequence includes remarkably good coverage of human telomere regions. For the 24 human chromosomes, we analysed 46 telomere ends in all. The telomeres of the sex chromosome pair X and Y recombine meiotically, so these four telomeres are treated as two (designated the Xp/Yp pseudoautosomal telomere and the Xq/Yq pseudoautosomal telomere). We could integrate the working draft sequence with 32 telomere regions captured by half-YAC clones (blue dots). Of these 32 regions, 18 have working draft sequence coverage that includes DNA less than 50 kilobases (kb) from the telomere; for five of them, the sequence extends to the terminal (TTAGGG) n sequences^{10–13} (see Supplementary Information). Although we were unable to capture two telomeres (5p and 20q) in half-YAC clones, we identified these regions in subtelomeric repeat-containing BAC clones and used them to connect to the working draft sequence (green dots).

We were unable to connect 7 of the remaining 12 telomere ends to the working draft, either because the working draft sequence does not yet extend into these regions (2q, 7p, 17p, 17q and Xp/Yp) or because unambiguous identification of overlapping working draft sequence was prevented by repeat sequences in the telomere clones (19p and 19q). BAC or cosmid clones connected to each of these seven telomeres were identified during construction of the fingerprint-based clone map of the human genome¹⁴ (<http://genome.wustl.edu/gsc/human/Mapping/index.shtml>) and this is likely to facilitate the future integration of these telomeres into the working draft (see Methods). The five acrocentric chromosomes (13, 14, 15, 21 and 22) contain heterochromatic short arms comprising repeated DNA. We did not analyse these five short-arm telomeres (black rectangles) because their sequences were unstable in

both yeast and bacteria, rendering them difficult to clone and characterize.

We have made available a detailed summary of the mapping experiments integrating telomeres with the working draft sequence, including specific telomere reagents, working draft contig designations, individual BAC clone accessions and accession numbers for our half-YAC-derived sequences (see Supplementary Information). For some chromosome ends, such as that of 11p (Fig. 2), we could precisely estimate the distance between the end of the working draft sequence and the telomere. However, this was not the case for many telomeric regions because much of the working draft sequence is still in small, unordered pieces.

As part of this study, around 1.1 megabases (Mb) of half-YAC-derived DNA sequence was acquired from cosmid end sequencing as well as from draft and finished sequencing of some subtelomeric cosmids (see Supplementary Information). In addition to defining the overlap relationship between the working draft sequence and the telomere clones, the half-YAC-derived sequences were useful for sampling the subtelomeric regions not yet included in the working draft sequence but captured in the half-YAC clones. Preliminary analysis of the half-YAC-derived sequences and the regions of the working draft sequence that overlapped with the half-YACs revealed several interesting features.

The sizes of subtelomeric repeat regions adjacent to the terminal (TTAGGG) n varied widely among individual telomeres, from 8 kb at the 7q telomere to ~300 kb at the 8p telomere. Large variations in subtelomeric repeat content have been detected near at least 18 telomeres^{8,15,16}. Nonetheless, the scale of human genomic subtelomeric repeat content is now well defined, and it is clear that a significant part of the subtelomeric repeat region of the human genome is present in the working draft sequence.

Large subtelomeric repeat regions can cause false linkages in the

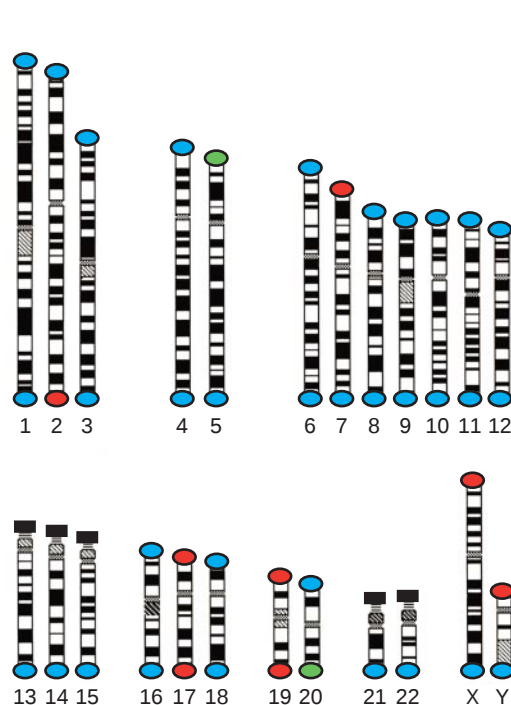


Figure 1 Summary of integration of telomeric DNA with working draft sequence. The two human pseudoautosomal telomere pairs (Xp/Yp and Xq/Yq) each recombine meiotically, so each pair is treated as a single telomere. Blue: the working draft sequence extends into these 32 telomere regions defined by half-YAC clones. Green: the working draft sequence ends within these subtelomeric repeat regions, but the distance to the molecular telomere is not known. Red: the telomeric DNA has not yet been integrated with working draft sequence. Telomere clones for the five acrocentric short-arm regions (black rectangles) have not been characterized.

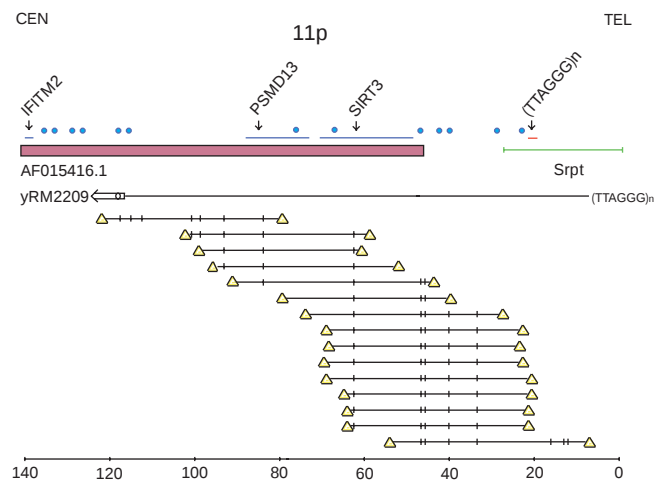


Figure 2 Connecting the 11p telomere to the working draft sequence. The end of the working draft sequence is represented by fragment AF015416 (magenta rectangle; a constituent of working draft contig 18272). The half-YAC clone yRM2209 (black) is represented below. Cosmid subclones were derived from the half-YAC clone and the contig of these subclones, which encompasses most of the half-YAC clone, was aligned by *EcoRI* restriction sites (small vertical marks). The cosmid end sequences (yellow triangles) were screened to the working draft sequence to orient the clones relative to the 11p telomere. This physical mapping enabled us to determine the size and features of the 45-kb telomeric region missing from the working draft sequence. Features include a 25-kb region of subtelomeric repeat DNA (green), an internal (TTAGGG) n telomere repeat sequence (red) and four Unigene clusters of ESTs (blue dots). Within the 140-kb region extending from the telomere, we determined the location of three known genes, *IFITM2* (an interferon-induced transmembrane protein), *PSMD13* (a component of the proteasome 26S subunit) and *SIRT3* (a Sir-2-like histone deacetylase implicated in telomere maintenance), as well as the positions of unigene clusters of ESTs that match 11p tel sequences and are distinct from known genes (blue dots).

BAC map and misassembly of working draft sequence. Large stretches of low-copy repeat DNA from subtelomeric repeat regions also localize to some pericentric chromosome regions, to the short-arm heterochromatin of acrocentric chromosomes and to a few loci in internal regions of chromosomes (for example, 1q42, 2q31, 4q28, 12p12 and Yq11.2). In previous iterations of the BAC map there were many instances of incorrect merges of subtelomeric repeat-containing BACs. To help identify these potentially problematic regions of the BAC map and working draft sequence, we have catalogued individual BAC clones containing segments of similarity with subtelomeric repeat regions (<http://www.wistar.upenn.edu/Riethman>). Inconsistencies between the current version of the BAC accession map (http://genome.wustl.edu:8021/pub/gsc1/fpc_files/freeze_2000_10_07/MAP/) and our telomere mapping studies are indicated in the Supplementary Information.

The abundance of low-copy repeat regions near telomeres is likely to make whole-genome shotgun assembly of subtelomeric regions virtually impossible. Indeed, previously characterized *Drosophila* subtelomeric repeat sequences are absent from its genome sequence¹⁷. By contrast, the entire sequence of yeast telomeric and subtelomeric regions was acquired using the half-YAC cloning strategy employed here¹⁸.

Internal telomere-like sequences, each consisting of around 50–250 base pairs (bp) of a mixture of perfect and imperfect copies of (TTAGGG)*n*¹¹, were present in all subtelomeric repeat regions analysed. For example, multiple copies of internal telomere-like sequences were present in widely spaced parts of the 100-kb 18p subtelomeric repeat region, and were present in both orientations relative to the telomere. It is interesting to speculate that packaging of subtelomeric chromatin might involve interactions between the terminal (TTAGGG)*n* repeats and these internal telomere sequences. The TRF1 protein, which binds to (TTAGGG)*n* *in vivo* and can bind sequences corresponding to the short internal repeats *in vitro*¹⁹, would be a good candidate for mediating such interactions.

Preliminary analysis of the potential gene content of the subtelomeric regions encompassed by the half-YAC-derived sequences and the overlapping portions of the subtelomeric working draft sequence was done by searching for sequence matches between the genomic DNA sequences and potential gene-derived complementary DNA and expressed sequence tag (EST) sequences in GenBank (<http://www.wistar.upenn.edu/Riethman>). Even this preliminary analysis reveals two features of subtelomeric regions. First, there are many sequence matches with genes and ESTs in most subtelomeric regions. We detected about 500 matches to transcripts identified by either a full-length cDNA or by a unigene cluster of expressed sequences in the 40 telomere regions analysed; 62 of these were found from half-YAC sequences mapping distal to the working draft sequence. Second, many of the genes and potential genes identified by sequence matches are members of gene families with many pseudogene copies. The sequence matches included around 100 known genes, both unique and members of gene families.

Human subtelomeric sequences have been proposed to serve as a buffer between the terminal (TTAGGG)*n* sequences, which are needed to protect chromosome ends from fusion and recombination, and vital internal chromosomal sequences¹⁵. However, the many expressed sequences throughout subtelomeric regions, extending almost to the molecular telomere, suggest that these regions may serve essential functions and are not simply dispensable junk DNA. □

Methods

We used a range of half-YAC-derived probes, including PCR- and cosmid subclone-derived probes and sequences (see Supplementary Information) and sets of collaboratively derived subtelomeric molecular and cytogenetic markers for specific telomeres^{20–22}, to connect specific cloned chromosome ends with flanking BAC contigs, either by DNA hybridization and PCR experiments or by computer-based matches (using BLAST²³ sequence alignment programs) of sequenced subtelomeric DNA with working draft sequence.

Single-copy probes from three of the seven telomeres not connected to working draft sequence could be used to identify BAC clones from an 11× coverage RP11 BAC library, although fewer clones than expected were identified (singleton BACs from the 2q and the 17p telomeres, and three BACs from the 7p telomere). Low-copy repeat sequences at the 19p, 19q and 17q telomeres complicated attempted BAC library screens for these chromosome ends, but independent experiments have identified PAC and BAC clones connected to the 17q telomere²² and a detailed physical map of chromosome 19 (<http://greengenes.llnl.gov/genome/>) exists to help guide closure of the 19p and 19q subtelomeric gaps, which occur in duplicated regions containing a family of zinc finger-encoding genes. The remaining telomere region (Xp/Yp) is encompassed by a 500-kb clone contig extending to within a few kb of the telomere²⁴.

Physical mapping experiments using a site-specific cleavage method (RARE cleavage^{8,25}) have been done for 21 telomeres to demonstrate co-linearity of the half-YAC insert DNA with the cognate telomere. In the absence of RARE cleavage data, the presence of subtelomeric repeats adjacent to terminal (TTAGGG)*n* sequences in all of the designated half-YAC clones is taken as strong evidence for proximity to the telomere; this has been borne out by the RARE cleavage experiments carried out so far.

Half-YAC clones containing chromosome-specific DNA were not recovered from four chromosome ends. BAC and cosmid clones identified by virtue of their subtelomeric repeat content form the initial basis for the telomere linkages to 5p, 20q, 19q and Xp/Yp. The BAC clones used to mark the 5p and 20q telomeres and the cosmid used to mark the 19q telomere each contain an internal telomere repeat sequence and subtelomeric repeat sequences, and localize to telomeric ends of the BAC map (5p, 20q) and the chromosome 19 physical map (<http://greengenes.llnl.gov/genome/>). On the basis of the known sequence organization of other telomeres, only additional subtelomeric repeat sequence is likely to reside distal to the subtelomeric repeat segments contained in these clones, although the possibility of single-copy DNA distal to them cannot be formally excluded at present. A cosmid clone mapped to the Xp/Yp pseudoautosomal telomere using *Bal31* exonuclease experiments²⁶ forms the telomeric boundary of a large cosmid contig²⁴ whose sequence is not yet available.

Received 15 November; accepted 18 December 2000.

- Blasco, M. A., Gasser, S. M. & Lingner, J. Telomeres and telomerase. *Genes Dev.* **13**, 2353–2359 (1999).
- deLange, T. & Jacks, T. For better or worse? Telomerase inhibition and cancer. *Cell* **98**, 273–275 (1999).
- McCulloch, R., Rudenko, G. & Borst, P. Gene conversions mediating antigenic variation in *Trypanosoma brucei* can occur on variant surface glycoprotein expression sites lacking 70-bp repeat sequences. *Mol. Cell Biol.* **17**, 833–843 (1997).
- Carlson, M., Celenza, J. L. & Eng, F. J. Evolution of the dispersed SUC gene family of *Saccharomyces* by rearrangements of chromosomal telomeres. *Mol. Cell Biol.* **5**, 2894–2902 (1985).
- Moyzis, R. K. *et al.* A highly conserved repetitive DNA sequence, (TTAGGG)*n*, present at the telomeres of human chromosomes. *Proc. Natl Acad. Sci. USA* **85**, 6622–6626 (1988).
- Riethman, H. C., Moyzis, R. K., Meyne, J., Burke, D. T. & Olson, M. V. Cloning human telomeric DNA fragments into *Saccharomyces cerevisiae* using a yeast-artificial-chromosome vector. *Proc. Natl Acad. Sci. USA* **86**, 6240–6244 (1989).
- Brown, W. R. *et al.* Structure and polymorphism of human telomere-associated DNA. *Cell* **63**, 119–132 (1990).
- Macina, R. A. *et al.* Molecular cloning and RARE cleavage mapping of human 2p, 6q, 8q, 12q, and 18q telomeres. *Genome Res.* **5**, 225–232 (1995).
- Doggett, N. A. *et al.* An integrated physical map of human chromosome 16. *Nature* **377** (Suppl.), 335–365 (1995).
- Flint, J. *et al.* The relationship between chromosome structure and function at a human telomeric region. *Nature Genet.* **15**, 252–257 (1997).
- Flint, J. *et al.* Sequence comparison of human and yeast telomeres identifies structurally distinct subtelomeric domains. *Hum. Mol. Genet.* **6**, 1305–1313 (1997).
- Ciccodicola, A. *et al.* Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. *Hum. Mol. Genet.* **9**, 395–401 (2000).
- Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–320 (2000).
- The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
- Wilkie, A. O. M. *et al.* Stable length polymorphism of up to 260 kb at the tip of the short arm of human chromosome 16. *Cell* **64**, 595–606 (1991).
- Trask, B. J. *et al.* Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet.* **7**, 13–26 (1998).
- Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–95 (2000).
- Louis, E. J. & Borts, R. A complete set of marked telomeres in *Saccharomyces cerevisiae* for physical mapping and cloning. *Genetics* **139**, 125–136 (1995).
- Bianchi, A. *et al.* TRF1 binds a bipartite telomeric site with extreme spatial flexibility. *EMBO J.* **18**, 5735–5744 (1999).
- Ning, Y. *et al.* A complete set of human telomeric probes and their clinical application. *Nature Genet.* **14**, 86–89 (1996).
- Rosenberg, M. *et al.* Characterization of short tandem repeats from thirty-one human telomeres. *Genome Res.* **7**, 917–923 (1997).
- Knight, S. J. L. *et al.* An optimized set of human telomere clones for studying telomere integrity and architecture. *Am. J. Hum. Genet.* **67**, 320–332 (2000).
- Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
- Gianfrancesco, F. A novel pseudoautosomal gene encoding a putative GTP-binding protein resides in the vicinity of the Xp/Yp telomere. *Hum. Mol. Genet.* **7**, 407–414 (1998).
- Riethman, H., Birren, B. & Gnirke, A. in *Genome Analysis: A Laboratory Manual*, Vol. 1, *Analyzing DNA* (eds Birren, B., Green, E., Klapholz, S., Meyers, R. & Roskams, J.) 83–248 (Cold Spring Harbor Laboratory Press, New York, 1997).
- Cooke, H. J. & Smith, B. A. Variability at the telomeres of the human X/Y pseudoautosomal region. *Cold Spring Harbor Symp. Quant. Biol.* **51**, 213–219 (1986).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank J. Finklestein, N. Atigapramoj, E. Dabagyan, S. Huang, A. Ambriz, A. Harxhi and K. Sutton for their contributions to this work, which was supported by NIH and DOE.

Correspondence and requests for materials should be addressed to H.C.R. (e-mail: Riethman@wistar.upenn.edu).

Comparison of human genetic and sequence-based physical maps

Adong Yu*, Chengfeng Zhao*, Ying Fan*, Wonhee Jang†, Andrew J. Mungall‡, Panos Deloukas‡, Anne Olsen§, Norman A. Doggett||, Nader Ghebranious*, Karl W. Broman¶ & James L. Weber*

* Center for Medical Genetics, Marshfield Medical Research Foundation, Marshfield, Wisconsin 54449, USA

† National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA

‡ The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

§ Joint Genome Institute, Walnut Creek, California 94598, USA

|| Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

¶ Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205-2179, USA

Recombination is the exchange of information between two homologous chromosomes during meiosis. The rate of recombination per nucleotide, which profoundly affects the evolution of chromosomal segments, is calculated by comparing genetic and physical maps. Human physical maps have been constructed using cytogenetics¹, overlapping DNA clones² and radiation hybrids³; but the ultimate and by far the most accurate physical map is the actual nucleotide sequence. The completion of the draft human genomic sequence⁴ provides us with the best opportunity yet to compare the genetic and physical maps. Here we describe our estimates of female, male and sex-average recombination rates for about 60% of the genome. Recombination rates varied greatly along each chromosome, from 0 to at least 9 centiMorgans per megabase (cM Mb^{-1}). Among several sequence and marker parameters tested, only relative marker position along the metacentric chromosomes in males correlated strongly with recombination rate. We identified several chromosomal regions up to 6 Mb in length with particularly low (deserts) or high (jungles) recombination rates. Linkage disequilibrium was much more common and extended for greater distances in the deserts than in the jungles.

All nucleated human cells contain two homologous copies of each chromosome, except for the sex chromosomes in males. During the formation of the sperm and egg cells, the number of each chromosome is reduced to one so that fertilization restores the normal diploid number in the next generation. The process of chromosome reduction, meiosis, is usually accompanied by exchange or recombination of DNA between the homologous parental chromosomes. Genetic maps, which are based on meiotic recombination, order and estimate distances between DNA sequences that vary between parental homologues (polymorphisms). The primary unit of distance along the genetic maps is the centiMorgan (cM), which is equivalent to 1% recombination.

The genetic maps used in our analysis were based upon the genotyping of 8,031 short tandem repeat polymorphisms (STRPs)

from Généthon, the University of Utah and the Cooperative Human Linkage Center in eight reference CEPH families⁵. Excluding the sex chromosomes, the maps cover about 4,250 cM in females and 2,730 cM in males. The genetic maps are relatively marker dense, with an average of 2–3 STRPs per cM, but are also relatively low resolution because only 184 meioses (92 in each sex) were analysed.

The physical maps used were all DNA sequence assemblies. For chromosomes 21 and 22, we used the finished, published sequences^{6,7}. For the other 20 autosomes and for the X chromosome, we used the public draft sequence assemblies, 5 September 2000 version (<http://genome.cse.ucsc.edu>)⁴. As we required relatively long stretches of sequence, we used only sequence assemblies that were over 1.5 Mb long (between terminal STRPs), contained more than three STRPs and had a marker order that agreed with published genetic and radiation hybrid maps. The amount of sequence used from each chromosome is shown in Fig. 1. Some chromosomes had much better coverage than others. We analysed 253 sequence assemblies ranging in length up to 70 Mb and spanning a total of 1,806 Mb (roughly 58% of the portion of the genome that is not highly repetitive). By far the most common reason for rejecting sequence assemblies was insufficient length; only seven assemblies were rejected for incompatible marker order.

Recombination rates varied greatly across the genome, from 0 to 8.8 cM Mb^{-1} (Table 1). Sex-average recombination rates (the average for males and females combined) did not vary as much as the sex-specific rates (for males and females considered separately) because male and female recombination rates at specific sites often differed substantially. We identified 19 recombination deserts up to 5 Mb in length with sex-average recombination rates below 0.3 cM Mb^{-1} , and 12 recombination jungles up to 6 Mb in length with sex-average recombination rates greater than 3.0 cM Mb^{-1} (see Supplementary Information). Wide variation in recombination rates across chromosomes has been observed previously for humans^{8–11} and for other eukaryotic species^{12–15}, and is clearly the rule rather than the exception.

In an effort to identify the basis of differences in recombination rates, we compared the rates to several marker and sequence parameters. These parameters included GC content, STRP informativeness, position of the marker relative to the centromeres and telomeres, density of runs of various short tandem repeats, especially (A)_n, (AC)_n, (AGAT)_n, (AAN)_n and (AAAN)_n sequences, and the density of various interspersed repetitive elements, including

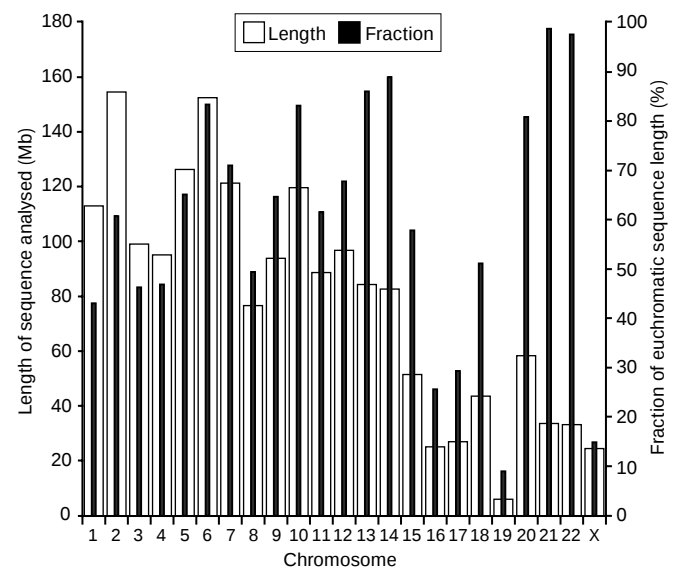


Figure 1 Sequence coverage for comparison of the genetic and physical maps. The total length of sequence used in the analysis (open bars) and the approximate percentage of the euchromatic sequence length (solid bars) are shown for each chromosome.

Table 1 Distribution of recombination rates

	Range		Mean $\pm$ s.d.	
Sex average	0.0–6.0		1.30 $\pm$ 0.80	
Male	0.0–7.9		0.92 $\pm$ 0.96	
Female	0.0–8.8		1.68 $\pm$ 1.07	
	0–0.5	0.5–1.5	1.5–3.0	>3.0
Sex average	12%	58%	26%	4%
Male	38%	45%	12%	5%
Female	10%	38%	42%	10%

Recombination rates given in cM Mb⁻¹. These data were derived from individual recombination rates at 4,088 STRPs.

Alu, L1, MIR (mammalian-wide interspersed repeats) and MER (medium reiterated repeats) sequences (Table 2).

With one exception, we found only weak correlations between the parameters and recombination rates. As controls, we found the expected negative correlation between short interspersed nuclear element (SINE) and long interspersed nuclear element (LINE) densities¹⁶, and the positive correlation between SINE density and (A)_n density. In agreement with ref. 17, we found no correlation between STRP heterozygosity and recombination rate, despite reports of positive correlation of nucleotide (sequence) diversity values with recombination rates^{18,19}. However, STRP heterozygosities are probably much more dependent upon relatively high mutation rates than selection and are therefore likely to be poor measures of nucleotide diversity. Similarly, we found only weak correlation between (AC)_n density ($n \geq 11$ or 19) and recombination rates despite the report of such a correlation for chromosome 22 (ref. 20). GC content is of interest because the genome appears to be segmented into isochores of varying GC content and because GC content is strongly correlated with gene density²¹. We did confirm a positive relationship between recombination and GC content (ref. 22), but the correlation was weak. By far the strongest relationship detected was for the position of the markers along the metacentric chromosome arms in males. Male (but not female) recombination rates increased markedly near the telomeres.

Some important limitations apply to our comparison of human genetic and physical maps. First, the resolution of the genetic maps is modest, owing to the small number of meioses examined. This places relatively broad confidence intervals on the genetic map distances and similar broad confidence intervals on the recombination rates. Only sex-average recombination rates smaller than about 0.3 cM Mb⁻¹ and greater than about 2.5 cM Mb⁻¹ are statistically different from uniform recombination at the $P = 0.05$ significance level. Second, the draft sequences used in our analysis were often short, contained many gaps and still had some errors in marker order. When the finished sequences become available, additional

recombination deserts and jungles, for example, will undoubtedly be discovered. Third, there is mounting evidence for at least modest individual and possibly population variation in recombination rates^{5,23,24}. The genetic maps in our analysis were based on meiosis in only eight mothers and eight fathers, all or nearly all of European ancestry. Examination of a large sample of individuals and/or other populations might give different results. Finally, our analysis is only a long-range (megabase) analysis. We can reach no conclusions about recombination over short (kilobase) ranges. There is growing evidence for recombination hot spots no more than a few kilobases long^{13,25,26}. Megabase-sized chromosomal segments may turn out to be comprised of regions with little or no recombination separated by short recombination hot spots. Perhaps the primary difference between recombination deserts and jungles lies in the density and strength of recombination hot spots.

Despite the limitations, there is strong evidence that our results are reliable first estimates of human recombination rates. Genetic maps based on 40 CEPH families show good agreement with the eight family maps (see, for example, ref. 8). Plots of the ratio of female to male recombination from the eight family data show maxima at the centromeres and minima at the telomeres for virtually all metacentric chromosomes⁵. The shapes produced by plotting centiMorgans against megabases obtained from the draft sequence assemblies for chromosomes 6 and 20 match closely those obtained using physical distances from restriction enzyme fingerprinting of overlapping genomic clones. Lengths of the draft sequence assemblies (17 July 2000 version) for chromosomes 21 and 22 matched the lengths of the finished sequences with only 0.1% error. And, probably most importantly, recombination deserts and jungles differ significantly in linkage disequilibrium (when two polymorphic alleles are not in random association).

The decay of linkage disequilibrium is expected to be much slower in recombination-poor than in recombination-rich regions. We tested this hypothesis by comparing linkage disequilibrium among pairs of STRPs within the recombination deserts and jungles. Although the power to detect linkage disequilibrium in genotyping data from only eight families is low, it was still found to be much higher for close pairs of markers in the deserts than in the jungles (Fig. 2). For marker pairs less than 0.5 Mb apart, 32% of pairs in the deserts showed significant linkage disequilibrium, as compared with only 7% in the jungles ($P = 0.001$).

In conclusion, our work shows that recombination rates vary greatly across the human genome, by at least two orders of magnitude. Linkage disequilibrium will generally extend over longer distances in regions with low recombination. Mapping genes responsible for traits and diseases by association studies will be easier and require a lower density of polymorphisms in regions of

Table 2 Correlations of recombination rates with sequence parameters

Parameter 1	Parameter 2	Relationship	R^2	P
SINE density	LINE density	Negative	0.20	<0.001
SINE density	(A) _n density	Positive	0.80	<0.001
Sex average recombination rate	STRP heterozygosity	None	0.00	0.49
Sex average recombination rate	GC content	Positive	0.05	<0.001
Sex average recombination rate	(AC) _n density ($n \geq 11$)	Positive	0.001	0.03
Sex average recombination rate	(A) _n density	Positive	0.02	<0.001
Sex average recombination rate	SINE density	Positive	0.01	<0.001
Sex average recombination rate	LINE density	Negative	0.02	<0.001
Male recombination rate	Chromosome position (metacentrics)	Positive towards telomeres	0.18	<0.001

Correlation was evaluated by linear regression analysis. Relationship indicates the sign of the slope of the best fitting line. R^2 is a measure of the fit of the points to the line. R^2 can vary from 0 to 1.0 with 1.0 indicating a perfect fit. P is the probability of obtaining an R^2 value as large as observed given no correlation. For the original plots from which these parameters were derived, see Supplementary Information.

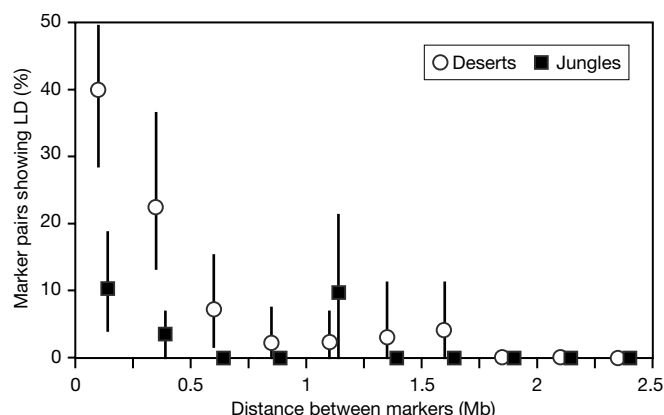


Figure 2 Linkage disequilibrium (LD) among pairs of STRPs within autosomal recombination deserts and jungles. Deserts and jungles are listed in the Supplementary Information. Marker pairs were binned into 0.25-Mb spacing intervals.

low recombination. Nucleotide and haplotype diversity will also probably parallel recombination rates. Although our baseline long-range recombination rates will be useful, they should be recalculated when the human genomic sequences are finished and as higher resolution genetic maps become available. In the more distant future, genotyping greater numbers of reference families at much higher polymorphism densities will lead to short-range maps of recombination hot spots. □

Methods

Connection of genetic and physical maps

We used short, single-pass genomic sequences and/or PCR primer sequences for STRPs to identify draft or finished bacterial artificial chromosome (BAC) or cosmid sequences within GenBank that encompass the STRPs using BLAST²⁷ and ePCR²⁸. Blast criteria were score (bits) > 200, expect (E) value < e⁻⁵⁰, and ratio of matched bases to marker sequence length > 85%. ePCR criteria were no more than one base mismatch in each primer and size of PCR product within allele size range for the STRP. About 75% of the STRPs were connected to the long genomic sequences. The reasons for failure of the remaining 25% are not fully understood, but include absence of the corresponding sequence in GenBank and poor quality of the STRP sequences. As the genetic maps are marker rich, the absence of 25% was not a serious limitation. Tables of STRPs with GenBank sequence accession numbers for encompassing BACs, genetic map positions and recombination rates are available from the Marshfield web site.

Determination of recombination rates

For each sequence assembly we built new female, male and sex-average genetic maps, using the marker order provided by the assemblies and using the genotyping data from the eight CEPH reference families⁵. We fitted cubic splines to plots of genetic versus physical distance, and from these curves we obtained recombination rates as first derivatives¹⁵. The statistical significance of the recombination rates was estimated by computer simulation of 1,000 iterations of recombination within each interval between markers, assuming a constant level of recombination across the genome for each sex. The constant levels of recombination were taken as the total genetic lengths of all the assemblies analysed divided by the total physical lengths of these assemblies.

Computation of marker and sequence parameters

We calculated STRP heterozygosities using genotypes of individuals within the eight CEPH families. We obtained STRP positions relative to centromeres and telomeres as the fractional sex-average genetic map distances from the centromeres to the telomeres (value of 0 for a STRP at the centromere and 1.0 for a STRP at the telomere)⁵. GC content and STR densities were obtained from programs written and tested at Marshfield²⁹. STR densities were measured as numbers of runs of non-interrupted repeats rather than total numbers of repeats. Minimum values of *n* for (A)_{*n*}, (AC)_{*n*}, (AGAT)_{*n*}, (AAN)_{*n*}, and (AAAN)_{*n*} sequences were 12, 11 or 19 ((AC)_{*n*}), 5, 7 and 5, respectively. We obtained interspersed repetitive element densities using the program Repeat Masker (<http://ftp.genome.washington.edu/RM/RepeatMasker.html>). SINEs and LINEs were defined by Repeat Masker and consist primarily of Alu and L1 elements, respectively. We computed all DNA sequence parameters over 250-kb windows centred about each STRP. For markers ≤ 125 kb from the ends of the sequence assemblies, we defined the window as the 125 kb of proximal sequence plus all available distal sequence. Unknown bases in the sequence assemblies were excluded from analysis. All parameters were corrected for reduced window size owing to unknown bases or proximity to ends.

Measurement of linkage disequilibrium

Recombination deserts and jungles were selected as those chromosomal regions with sex-average recombination rates of <0.3 or >3.0, respectively. We measured linkage disequilibrium for all pairs of STRPs within the deserts (449 pairs) and jungles (467 pairs) using Fisher's exact test³⁰. Only disequilibrium results that were significant at *P* ≤ 0.01 were plotted in Fig. 2. An overall *P*-value was obtained by a permutation test treating the regions as units in order to account for the dependence between marker pairs within a region.

Received 27 October; accepted 8 December 2000.

1. The BAC Resource Consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**, 953–958 (2001).
2. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
3. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
4. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
5. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
6. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
7. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
8. Fain, P. R., Kort, E. N., Yousry, C., James, M. R. & Litt, M. A high resolution CEPH crossover mapping panel and integrated map of chromosome 11. *Hum. Mol. Genet.* **5**, 1631–1636 (1996).
9. Bouffard, G. G. *et al.* A physical map of human chromosome 7: An integrated YAC contig map with average STS spacing of 79 kb. *Genome Res.* **7**, 673–692 (1997).
10. Nagaraja, R. *et al.* X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. *Genome Res.* **7**, 210–222 (1997).

11. Mohrenweiser, H. W., Tsujimoto, S., Gordon, L. & Olsen, A. S. Regions of sex-specific hypo- and hyper-recombination identified through integration of 180 genetic markers into the metric physical map of human chromosome 19. *Genomics* **47**, 153–162 (1998).
12. Nicolas, A. Relationship between transcription and initiation of meiotic recombination: toward chromatin accessibility. *Proc. Natl Acad. Sci. USA* **95**, 87–89 (1998).
13. Wahls, W. P. Meiotic recombination hotspots: shaping the genome and insights into hypervariable minisatellite DNA change. *Curr. Top. Dev. Biol.* **37**, 37–75, (1998).
14. Faris, J. D., Haen, K. M. & Gill, B. S. Saturation mapping of a gene-rich recombination hot spot region in wheat. *Genetics* **154**, 823–835 (2000).
15. Kliman, R. M. & Hey, J. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol. Biol. Evol.* **10**, 1239–1258 (1993).
16. Chen, T. L. & Manucladis, L. SINEs and LINEs cluster in distinct DNA fragments of Giemsa band size. *Chromosoma* **98**, 309–316 (1989).
17. Payseur, B. A. & Nachman, M. W. Microsatellite variation and recombination rate in the human genome. *Genetics* **156**, 1285–1298 (2000).
18. Nachman, M. W., Bauer, V. L., Crowell, S. L. & Aquadro, C. F. DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**, 1133–1141 (1998).
19. Nachman, M. W. & Crowell, S. L. Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, *Dmd*, in humans. *Genetics* **155**, 1855–1864 (2000).
20. Majewski, J. & Ott, J. GT repeats are associated with recombination on human chromosome 22. *Genome Res.* **10**, 1108–1114 (2000).
21. Bernardi, G. Isochores and the evolutionary genomics of vertebrates. *Gene* **241**, 3–17 (2000).
22. Eisenbarth, L., Vogel, G., Krone, W., Vogel, W. & Assum, G. An isochore transition in the NF1 gene region coincides with a switch in the extent of linkage disequilibrium. *Am. J. Hum. Genet.* **67**, 873–880 (2000).
23. Yu, J. *et al.* Individual variation in recombination among human males. *Am. J. Hum. Genet.* **59**, 1186–1192 (1996).
24. Lien, S., Szzyda, J., Schechinger, B., Rappold, G. & Arnheim, N. Evidence for heterogeneity in recombination in the human pseudoautosomal region: High resolution analysis by sperm typing and radiation-hybrid mapping. *Am. J. Hum. Genet.* **66**, 557–566 (2000).
25. Carrington, M. Recombination within the human MHC. *Immunol. Rev.* **167**, 245–256 (1999).
26. Jeffreys, A. J., Ritchie, A. & Neumann, R. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hot spot. *Hum. Mol. Genet.* **9**, 725–733 (2000).
27. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389 (1997).
28. Schuler, G. D. Sequence mapping by electronic PCR. *Genome Res.* **7**, 541–550 (1997).
29. Zhao, C., Heil, J., & Weber, J. L. A genome wide portrait of short tandem repeats. *Am. J. Hum. Genet.* **65**, (Suppl.) A102 (1999).
30. Huttley, G. A., Smith, M. W., Carrington, M. & O'Brien, S. J. A scan for linkage disequilibrium across the human genome. *Genetics* **152**, 1711–1722 (1999).

Supplementary information is available from Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of *Nature*, and from the Marshfield Web site (<http://research.marshfieldclinic.org/genetics>).

Acknowledgements

This work was supported by contracts from the US National Institutes of Health and Department of Energy. Assistance was also provided by the chromosome 6 and 20 project groups at the Sanger Centre, supported by the Wellcome Trust.

Correspondence and requests for materials should be addressed to J.L.W. (e-mail: weberj@cmg.mfldclin.edu).

Integration of cytogenetic landmarks into the draft sequence of the human genome

The BAC Resource Consortium*

* *Authorship of this paper should be cited using the names of authors that appear at the end.*

We have placed 7,600 cytogenetically defined landmarks on the draft sequence of the human genome to help with the characterization of genes altered by gross chromosomal aberrations that cause human disease. The landmarks are large-insert clones mapped to chromosome bands by fluorescence *in situ* hybridization. Each clone contains a sequence tag that is positioned on the genomic sequence. This genome-wide set of sequence-anchored clones allows structural and functional analyses of the genome. This resource represents the first comprehensive integration of cytogenetic, radiation hybrid, linkage and sequence maps of the

human genome; provides an independent validation of the sequence map^{1,2} and framework for contig order and orientation; surveys the genome for large-scale duplications, which are likely to require special attention during sequence assembly; and allows a stringent assessment of sequence differences between the dark and light bands of chromosomes. It also provides insight into large-scale chromatin structure and the evolution of chromosomes and gene families and will accelerate our understanding of the molecular bases of human disease and cancer.

With the draft of the human genome available², scientists can conduct global analyses of its gene content, structure, function and variation. One important challenge is to define the genetic contribution to human diseases. For many developmental disorders, inherited conditions and cancers, gross chromosomal aberrations provide clues to the locations of the causative molecular defects. These aberrations are visible as alterations in chromosomal banding patterns³ or in the number or relative positions of DNA sequences labelled by fluorescence *in situ* hybridization (FISH)⁴. Although tracing gross abnormalities to the level of DNA sequence⁵ has revealed the genetic causes of many diseases, molecular characterization of chromosomal aberrations has lagged far behind their discovery⁶. To proceed from cytogenetic observation to gene discovery and mechanistic explanation, scientists will need access to a resource of experimental reagents that effectively integrates the cytogenetic and sequence maps of the human genome.

We describe here the results of a concerted effort to assemble such a genome-wide resource of well mapped, large-insert DNA clones. Each clone has been localized directly to chromosomal band(s) by FISH (Fig. 1a) and assigned one or more unique sequence tags, which can anchor the clone to the emerging draft sequence. We used complementary strategies to amass the current set of 8,877 clones.

The set, which consists primarily of bacterial artificial chromosome (BAC) clones, includes clones targeted to contain sequence-tagged sites (STSs) ordered along the genome by genetic linkage or radiation hybrid mapping (for well ordered and distributed coverage); clones randomly selected for end sequencing from the RPCI-11 library (for coverage of regions low in STSs); clones identified during intense mapping efforts that preceded sequencing of some chromosomes (for denser coverage); and clones suspected of being partially

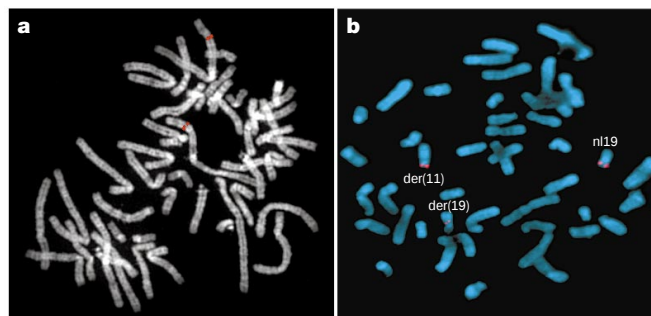


Figure 1 Cytogenetic analyses of sequence-integrated clones. **a**, Using FISH, fluorescent signals are observed at cytogenetic bands (grey) where fragments of a sequence-tagged BAC hybridize (red). **b**, Clones selected on the basis of band location were used in FISH analyses to map the breakpoint of a translocation involving chromosomes 11 and 19 in a patient with multiple congenital malformations and mental retardation (DGAP012, <http://dgap.harvard.edu>). Clone CTD-3193o13 spans the breakpoint on chromosome 19; red signal is split between the derivative chromosome 11 and derivative 19 chromosomes and is also present on the normal chromosome 19. The GTG-banded karyotype for this patient is 46,XY,t(11;19)(p11.2;p13.3).

Table 1 A cytogenetic resource of FISH-mapped, sequence-tagged clones

Chromosome	Number of FISH-mapped clones					Connections to the draft sequence			
	Type of sequence tag			Coverage		No. of clones anchored to draft sequence§	% discordant		
	Accession*	STS or gene	BAC end	Total†	Avg. density‡		Chrom.	Site	% concordant
1	868	318	355	1,248	4.9	1,127	2	2	95
2	43	180	189	297	1.2	241	3	2	95
3	128	222	178	308	1.5	233	5	6	90
4	42	253	227	341	1.7	275	7	1	92
5	35	237	168	296	1.6	255	3	4	93
6	653	212	176	909	5.1	801	3	1	96
7	25	254	151	324	1.9	274	2	1	96
8	31	181	161	245	1.6	203	5	3	92
9	208	169	252	384	2.7	324	4	2	94
10	191	302	288	454	3.2	382	4	4	93
11	119	243	225	378	2.7	324	6	2	92
12	109	251	178	304	2.2	266	7	2	91
13	182	101	175	278	2.4	252	3	1	96
14	48	167	167	222	2.1	196	3	2	96
15	109	117	154	224	2.2	189	5	2	94
16	72	237	196	267	2.8	222	4	2	95
17	21	71	77	119	1.3	93	10	1	89
18	9	73	76	105	1.3	86	2	0	98
19	7	55	49	76	1.1	56	14	0	86
20	228	107	112	388	5.6	333	1	1	98
21	4	64	52	85	1.8	72	1	0	99
22	217	123	108	343	6.5	303	3	1	96
X	641	274	150	872	5.5	782	2	2	96
Y	7	13	15	17	0.3	14	7	0	93
Subtotal	3,997	4,224	3,879	8,484	2.6	7,303	3.6	1.9	95
Multiple sites¶	209	100	220	393	n.a.	297	n.a.	n.a.	n.a.
Total	4,206	4,324	4,099	8,877	n.a.	7,600			

All clones are associated with a sequence-tag; localized directly to cytogenetic bands by FISH; BACs, P1, or PACs; archived as single-colony-purified stocks; and publicly available.

n.d., not done. n.a., not applicable.

* Clones whose draft or finished sequence is deposited in GenBank.

† Total is less than sum of preceding three columns because some clones have >1 type of sequence tag.

‡ In clones per Mb, that is, number of FISH-mapped clones/chromosome size in Mb²⁰.

§ Sequence tags of 8,325 single-site and 352 multisite clones were used to search the 7 October 2000 draft. Clones whose sole tag consisted of a Unigene accession (Hs.) and some multisite clones have not yet been evaluated.

|| Discordant site refers to clones mapped by FISH to location >1 band away from, but on same chromosome as, neighbours on draft.

¶ Because most clones in the resource were not selected randomly, fraction of multi-site clones does not accurately reflect frequency of low-copy duplications in the genome.

duplicated at more than one location in the genome (to flag regions of the genome that might complicate sequence assembly⁷). The molecular signatures are STSs (many corresponding to genes or expressed sequence tags (ESTs)), BAC end sequences, or the actual draft or final sequence of the clone (Table 1). Earlier publications have described genome-wide and chromosome-specific subsets of this collection^{8–12}.

Each clone is publicly available as single-colony-purified bacterial stocks and is ready for distribution. Each clone can each be obtained from one of three stock centres by e-mail: mapped-clones@mail.cho.org, libraries@resgen.com and clonerequest@sanger.ac.uk. The website <http://www.ncbi.nlm.nih.gov/genome/cyto> provides information about all clones in this collection, including how to obtain each clone. (Additional information can be obtained at the websites listed in Supplementary Information 1).

The 8,877 clones provide excellent coverage of the human genome (Table 1), with at least one clone on average per megabase (Mb) for 23 of the 24 chromosomes. Clone density ranges from greater than ~5 clones per Mb for chromosomes 1, 6, 20, 22 and X to about 0.3 clones per Mb for chromosome Y.

Our study provides an assessment of the representation of the human genome in the RPCI-11 BAC library¹³, which serves as the intermediate template for most sequencing efforts² and the foundation of genome-wide contig assembly by fingerprint analyses¹. We

randomly selected 1,243 clones from this library for FISH analysis. The number of clones assigned to each chromosome correlated well with chromosome size, with no significant bias in the distribution of clones between Giemsa (G)-dark and G-light bands of chromosomes (see Supplementary Information 2 and 3).

Cytogenetic mapping is one of several methods that can produce a framework of ordered clones upon which the human sequence can be assembled. The resource provides an opportunity to cross-check these critical framework maps, because over 3,300 FISH-mapped clones have STSs that reference the radiation hybrid¹⁴ or linkage maps^{15,16}. Overall, the concordance between cytogenetic map order and marker order established by radiation hybrid and linkage mapping is very high for clones with single cytogenetic locations (94–98%, depending on the map; Table 2). Significant discrepancies were observed for only around 140 of these clones and are probably due to errors in clone tracking. Integration of cytogenetic and linkage maps also aids efforts to map disease genes. The location of the cytogenetic abnormality in one patient can guide the choice of polymorphic markers to assess linkage in other families that have similar phenotypes, but no visible chromosomal aberrations.

At present, 7,303 clones that map to single cytogenetic locations are positioned by their sequence tags on the draft sequence assembly of 7 October 2000 (Table 1). The fraction of clones located on the draft sequence ranges from 76% to 91% across different

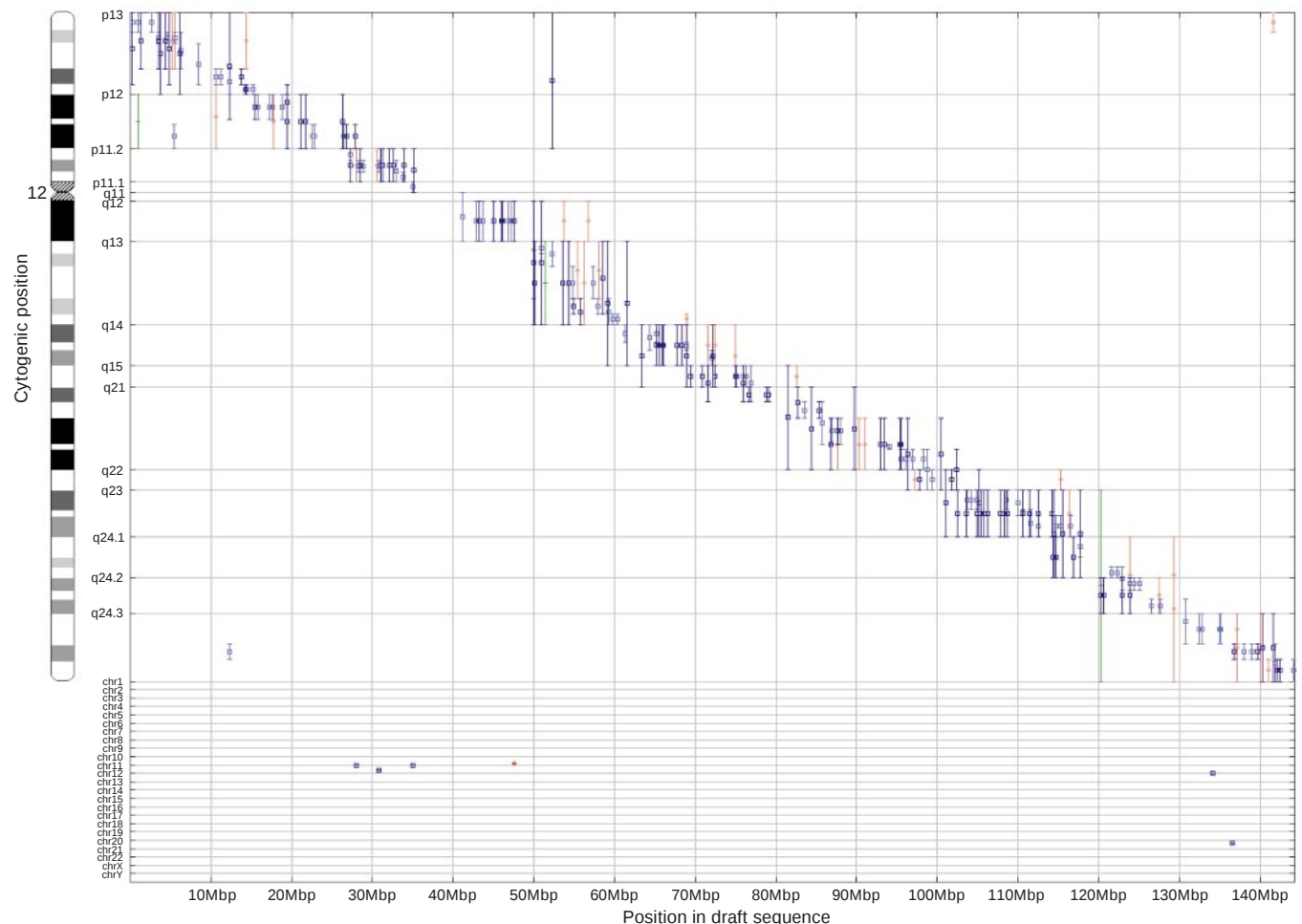


Figure 2 The correspondence between cytogenetic location and position on the 7 October 2000 draft sequence for chromosome 12. The band location of each clone is indicated by a range on the y-axis. Clones mapping to chromosomes other than 12 are indicated at the bottom. Colours differentiate assignments made in different laboratories. Each clone is anchored on the draft sequence by one or more sequence tags. Plots for the other chromosomes and the 5 September, 2000 assembly can be found at <http://>

genome.ucsc.edu/goldenPath/mapPlots/. Genome browsers that assist researchers in navigating from cytogenetic location to other maps and detailed, annotated sequence information are available at http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch (NCBI Mapviewer, which includes chromosomal aberrations associated with cancer and inherited disorders), <http://www.ensembl.org/> and <http://genome.ucsc.edu>.

Table 2 Clones connecting the cytogenetic map and other maps of the human genome

Map type	Version	Number	% concordant	% discordant	
				Chrom.	Site
Genetic	Genethon	1,686	98	1.4	0.7
	Marshfield	1,757	98	1.4	0.9
Radiation hybrid	GM99-GB4	1,433	98	1.3	1.2
	GM99-G3	1,654	96	2.5	1.4
	TNG	908	94	2.9	2.5
	5 Sept. 2000	7,364	94	3.7	2.2
Draft sequence	7 Oct. 2000	7,303	95	3.6	1.9

Many clones have markers positioned on more than one map. Only clones assigned to single chromosome locations by FISH are considered above. An additional 91 clones that map by FISH to more than one location contain STSs placed on other maps. STSs are by definition unique, single-copy markers, so each is assigned to a single genomic location by other mapping approaches. In 88% of these 91 cases, the STS location corresponds to one of the FISH-detected locations.

chromosomes (see Supplementary Information 4). We expect these percentages to rise as more sequence is merged into the draft and algorithms for locating tags are refined.

The connections between the cytogenetic map and the draft sequence are well distributed across the genome, and the correspondence in position on the two maps is excellent for these 7,303 clones (Fig. 2 shows chromosome 12 as an example). Of the 943 contigs of overlapping clones in the 7 October 2000 draft sequence, 660 are connected to the cytogenetic map by at least one clone, and 531 by two or more clones. Thus, many contigs can be oriented on the chromosome on the basis of FISH results of constituent clones. Relatively few discrepancies between cytogenetic location and position in the draft sequence are apparent at this level of resolution (~5% of the clones map either to other chromosomes or more than one band away from the expected position; Table 1). We found only eight locations where the cytogenetic data indicated that portions of the sequence were misplaced within an earlier draft assembly (5 September 2000). The sequencing centres used these cytogenetic findings to locate errors in the assembly and produce the later draft of improved quality (Table 2).

FISH analyses of this clone collection reveal abundant paralogous relationships among sites dispersed across the human genome. Of 1,243 clones randomly selected from the RPCI-11 library, 5.4% hybridize to more than one chromosomal location (see Supplementary Information 3). The entire collection includes 393 clones that together identify over 150 bands containing at least one segment with significant homology to one or more (up to 25) other sites in the genome (see Supplementary Information 5). These data provide clues to duplications and exchanges that have occurred within and between chromosomes. Among the 393 clones, 111 contain blocks duplicated within the same chromosome; 282 hybridize to more than one chromosome. Paralogous relationships involving pericentromeric and subtelomeric regions of multiple chromosomes are particularly frequent and complex. Clones in the collection also identify low-copy duplications specific to chromosomes 1, 7, 11 and 16, the pseudoautosomal regions of X and Y, and sites of the olfactory receptor gene family¹⁷. Many previously undescribed patterns were also observed; some were confirmed with two or more clones, but others require further study to verify that they reflect true duplications.

Many of these duplications are functionally significant, as some have generated multigene families, and some are potential sites of recombination events, which can result in chromosome abnormalities. The cytogenetic data should greatly facilitate analyses of these regions, which are likely to pose challenges to sequence assembly. The sequence tags of 84% of the clones that hybridize to more than one site were placed in the 7 October 2000 draft assembly, and the location(s) were roughly consistent with at least one FISH observation for 88% of these clones. Collectively, the multisite clones highlight regions that are more likely to become entangled with other regions of the genome during sequence assembly than clones with single FISH locations. Indeed, global BLAST analyses show that

regions encompassing sequence tags of multi-site clones (either the sequence of the FISH-mapped clone or a surrogate clone from the assembly) contain blocks of homology found at an average of around 3.9 chromosomal locations (compared to around 1.3 for the regions underlying clones with single FISH signals). The regions observed by FISH and revealed through homology searches are not fully congruent, however (not shown). These findings indicate that both FISH and sequence analyses may underestimate large-scale duplications and that these complex, inter-related regions of the genome will require special attention during the finishing stages of genome sequencing.

The extensive integration of cytogenetic and primary sequence data gives investigators access to fine-structure information—including details on predicted genes—for cytogenetic locations of interest. Tools such as NCBI's MapViewer and the UCSC and ENSEMBL genome browsers (see Fig. 2 for URLs) allow researchers to navigate readily between chromosomal location and annotated sequence.

This integration provides insight into the sequence differences underlying cytogenetic banding patterns. Sequence analyses of 200-kilobase (kb) regions surrounding the sequence tags of 338 clones mapped with the finest band resolution reveal more striking differences in the base-pair composition between Giemsa-positive and -negative bands than were predicted from earlier studies¹⁸. These clones were mapped with high precision to 850-level bands of varying staining intensity¹⁹ on seven chromosomes. The AT content of 58 of the 59 clones in the darkest G-bands exceeds the genome-wide average of 0.59 (mean 0.63), whereas the AT content of only 22 of the 143 clones in G-negative bands is higher than average (mean 0.55; $\chi^2 = 43$, $P < 0.005$). These data confirm that dark G-bands are more AT-rich than G-negative bands.

The utility of a sequence-integrated cytogenetic resource is illustrated by two examples. In the first, clones are applied in conventional FISH assays to rapidly narrow the search for candidate genes disrupted or deregulated by translocations causing developmental disorders. The process is expedited by selection of clones assigned to the regions implicated by banding analyses. In a patient with multiple congenital malformations and mental retardation (DGAP012, <http://dgap.harvard.edu>), a breakpoint-spanning clone was identified (Fig. 1b). This clone spans a 170-kb interval containing the gene for MKK7, a human mitogen-activated protein kinase, and a novel sequence with homology to the *tre-2* oncogene, both plausible candidate genes. More typically, breakpoints will be mapped to an interval between neighbouring clones. For example, a translocation implicated in mental retardation in another patient maps to an interval containing at least 12 genes, including protocadherin 8, a promising candidate given its exclusive expression in fetal and adult brain²⁰.

In the second example, an array of around 2,000 BAC clones from the collection is used to perform a genome-wide scan for segmental aneuploidy by comparative genomic hybridization (CGH) (Fig. 3 and A. Snijders *et al.*, in preparation). The array format offers better sensitivity and resolution^{21,22} than metaphase chromosomes, the traditional target for CGH²³, and, because the arrayed clones are integrated into the draft, copy-number abnormalities can be related directly to sequence information. To illustrate the power of array CGH, the ML-2 cancer cell line was 'karyotyped' using the array. Array CGH revealed relative copy-number losses on 1p, 6q, 11q and 20p and gains of 12, 13 and 20q (Fig. 3). Copy-number abnormalities on chromosomes 6, 11 and 20 were subsequently confirmed by FISH using clones predicted by array CGH to be included in the region of loss. Several of these alterations were noted in previous banding analyses (1p-, 6q-, 11q-, +12, +13q+)²⁴, but array CGH locates the breakpoints precisely relative to BACs that reference specific locations in the sequence.

More than 7,500 clones now link the cytogenetic map and sequence of the human genome. Application of these reagents in

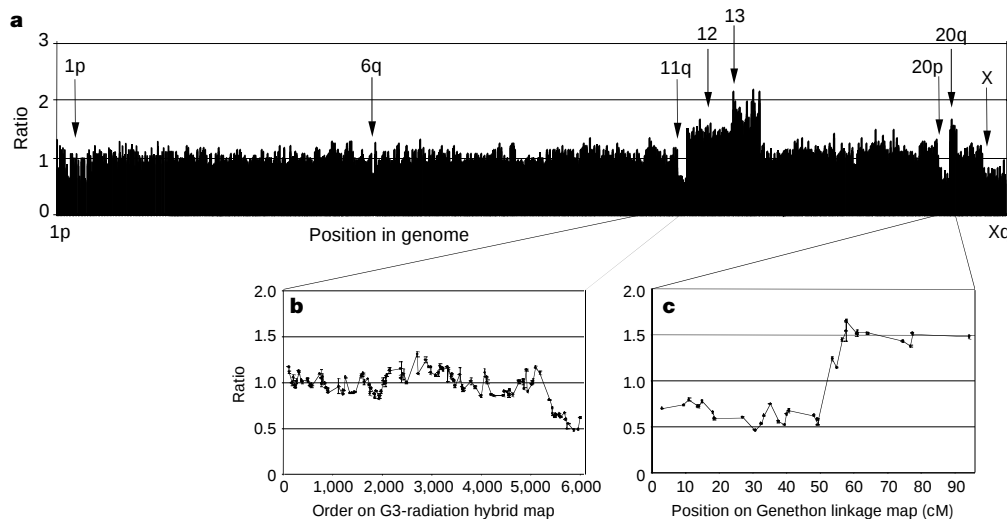


Figure 3 Copy-number analysis of myeloblastic leukaemia ML-2 cell line using CGH and a genome-wide array of around 2,000 BAC clones. The ML-2 cell line has acquired chromosomal abnormalities in addition to those present in the original tumour during long-term culture. CGH maps regions of abnormal copy number by comparing the relative efficiency with which test (Cy3-labelled ML-2 DNA) and reference (Cy5-labelled normal female DNA) hybridize to clones on the array. The array excludes clones that hybridize to multiple sites in the genome. **a**, Fluorescence ratios of Cy3 to Cy5 fluorescence for each

BAC normalized to the median ratio for all 2,000 clones on the array, ordered from 1pter to Xqter. Arrows, chromosomal regions showing significant copy number variations. The lower ratio on the X indicates expected ratio for mismatched sex of test and reference DNAs. Fluorescence ratios of clones on chromosomes 11 (**b**) and 20 (**c**) are shown with clones ordered according to position of their STSs on the G3 radiation hybrid or Genethon linkage maps, respectively.

combination with increasingly detailed knowledge of genes and other functional motifs in the human sequence will transform the process of identifying genes that are altered in cancer and other diseases. Ultimately, this resource will contribute to a better understanding of the organization of the cell nucleus, the compacting of DNA into mitotic chromosomes, and the basis of the chromosomal banding patterns that have been so valuable in uncovering the aetiology of human diseases. □

Methods

GenBank was screened for draft, finished or end sequences derived from clones in this collection. BACs were screened for STS content by a combination of hybridization and polymerase chain reaction (see refs 8, 25 and Supplementary Information for details). Sequence tags were located on the draft sequence by a combination of methods (see Supplementary Information and refs 26, 27). Sequence at these locations was compiled with the results of a genome-wide BLAST analysis (ref. 2 and J. A. Bailey and E. E. Eichler, in preparation) to identify paralogous regions of the genome (regions in the draft sequence containing ~20 kb of sequence that match sequence of the FISH-mapped clone or that of a surrogate clone from the assembly at ~90% identity in non-repeat-masked bases over each 1-kb segment), and these locations were translated into estimated band positions using a dynamic programming algorithm (T. S. Furey *et al.*, in preparation; and see Supplementary Information).

Details of FISH procedures are provided elsewhere^{4,28}. Only locations of unique or low-copy portions of the clone are identified, because high-copy interspersed repetitive sequences were suppressed by addition of unlabelled *Cot1* DNA. Replicate analyses indicate that the precision of FISH assignments to metaphase bands is roughly 5–10 Mb (1–1.5 band). A subset of 442 clones was ordered at very high (~2–3-Mb) resolution¹¹. FISH analyses were performed using DNA from the bacterial stock used for STS typing. Data that failed to replicate (for example, replicate FISH analyses of the same clone or different clones assigned the same marker) have been removed. Hybridization to arrays was carried out as described previously²⁹ and by Snijders *et al.* (in preparation).

Received 7 November 1999; accepted 20 December 2000.

1. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Caspersson, T. *et al.* Chemical differentiation along metaphase chromosomes. *Exp. Cell Res.* **49**, 219–222 (1968).
4. Trask, B. J. in *Genome Analysis: A Laboratory Manual* Vol. 4, 303–413 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1999).
5. Collins, F. S. Positional cloning moves from perdictional to traditional. *Nature Genet.* **9**, 347–350 (1995).
6. Mitelman, F. *Catalog of Chromosome Aberrations in Cancer* (Wiley, New York, 1998).
7. Eichler, E. E. Masquerading repeats: paralogous pitfalls of the human genome. *Genome Res.* **8**, 758–762 (1998).

8. Cheung, V. G. *et al.* A resource of mapped human bacterial artificial chromosome clones. *Genome Res.* **9**, 989–993 (1999).
9. Korenberg, J. R. *et al.* Human genome anatomy: BACs integrating the genetic and cytogenetic maps for bridging genome and biomedicine. *Genome Res.* **9**, 994–1001 (1999).
10. Leversha, M. A., Dunham, I. & Carter, N. P. A molecular cytogenetic clone resource for chromosome 22. *Chromosome Res.* **7**, 571–573 (1999).
11. Kirsch, I. R. *et al.* A systematic, high-resolution linkage of the cytogenetic and physical maps of the human genome. *Nature Genet.* **24**, 339–340 (2000).
12. Kirsch, I. R. & Ried, T. Integration of cytogenetic data with genome maps and available probes: Present status and future promise. *Semin. Hematol.* **37**, 420–428 (2000).
13. Osoegawa, K. *et al.* Bacterial artificial chromosome library for sequencing the human genome. *Genome Res.* (in the press).
14. Olivier, M. *et al.* A high resolution radiation hybrid map of the human genome draft sequence. *Science* (in the press).
15. Yu, A. *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953 (2001).
16. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
17. Trask, B. J. *et al.* Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Hum. Mol. Genet.* **7**, 2007–2020 (1998).
18. Saccone, S. *et al.* Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl. Acad. Sci. USA* **90**, 11929–11933 (1993).
19. Mitelman, F. (Ed.) *ISCN (1995): An International System for Human Cytogenetics Nomenclature* (S. Karger, Basel, 1995).
20. Strehl, S. *et al.* Characterization of two novel protocadherins (PCDH8 and PCDH9) localized on human chromosome 13 and mouse chromosome 14. *Genomics* **53**, 81–89 (1998).
21. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet.* **20**, 207–211 (1998).
22. Solinas-Toldo, S. *et al.* Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosom. Cancer* **20**, 399–407 (1997).
23. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818–821 (1992).
24. Ohyashiki, K., Ohyashiki, J. H. & Sandberg, A. V. Cytogenetic characterization of putative human myeloblastic leukemia cell lines (ML-1, -2, and -3): origin of the cells. *Cancer Res.* **46**, 3642–3647 (1986).
25. Morley, M. GenMapDB: A database of mapped human BAC clones. *Nucleic Acids Res.* **29**, 144–147 (2001).
26. Schuler, G. D. Electronic PCR: bridging the gap between genome mapping and genome sequencing. *Trends Biotechnol.* **16**, 456–459 (1998).
27. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214 (2000).
28. Korenberg, J. R., Chen, X. -N. Human cDNA mapping using a high-resolution R-banding technique and fluorescence in situ hybridization. *Cytogenet. Cell Genet.* **69**, 196–200 (1995).
29. Albertson, D. G. *et al.* Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nature Genet.* **25**, 144–146 (2000).
30. Trask, B. J., van den Engh, G., Mayall, B. & Gray, J. W. Chromosome heteromorphism quantified by high resolution bivariate flow karyotyping. *Am. J. Hum. Genet.* **45**, 738–752 (1989).

Supplementary information is available from Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank M. Arcaro, M. Bakis, J. Burdick, J. Chang, H.-C. Chen, S. Chiu, Y. Fan, C. Harris, L. Haley, R. Hosseini, J. Kent, M. A. Leversha, J. Martin, L.-T. Nguyen, P. Quinn, Y. H. Ramsey, T. Reppert, L. J. Rogers, J. Shreve, J. Stalica, M. Wang, T. Weber, A. M. Yavor, J. Young, K. Zatloukal, and members of the TIGR BAC Ends Team for assistance. This work was supported by grants from NIH (NCI, NHGRI, NIDCD and NICHD), US DOE, NSF, HHMI, PPG, Merck Genome Research Institute, Vysis, Inc., and start-up funds provided by Obstetrics and Gynecology at Brigham and Women's Hospital.

Correspondence should be addressed to B.J.T.
(e-mail: btrask@fhcrc.org).

The BAC Resource Consortium

V. G. Cheung^{1*}, N. Nowak^{2*}, W. Jang³, I. R. Kirsch⁴, S. Zhao⁵, X.-N. Chen⁶, T. S. Furey⁷, U.-J. Kim⁸, W.-L. Kuo⁹, M. Olivier¹⁰, J. Conroy², A. Kasprzyk¹¹, H. Massa¹², R. Yonescu⁴, S. Sait², C. Thoreen¹³, A. Snijders⁹, E. Lemyre¹⁴, J. A. Bailey¹⁵, A. Bruzel¹, W. D. Burrill¹¹, S. M. Clegg¹¹, S. Collins¹³, P. Dhami¹¹, C. Friedman¹², C. S. Han¹⁶, S. Herrick¹⁴, J. Lee⁸, A. H. Ligon¹⁴, S. Lowry¹⁷, M. Morley¹, S. Narasimhan¹, K. Osoegawa^{2,18}, Z. Peng¹⁷, I. Plajzer-Frick¹⁷, B. J. Quade¹⁴, D. Scott¹⁷, K. Sirotkin³, A. A. Thorpe¹¹, J. W. Gray⁹, J. Hudson¹⁹, D. Pinkel⁹, T. Ried⁴, L. Rowen²⁰, G. L. Shen-Ong⁴, R. L. Strausberg⁴, E. Birney¹¹, D. F. Callen²¹, J.-F. Cheng¹⁷, D. R. Cox¹⁰, N. A. Doggett¹⁶, N. P. Carter¹¹, E. E. Eichler¹⁵, D. Haussler²², J. R. Korenberg⁶, C. C. Morton¹⁴, D. Albertson⁹, G. Schuler³, P. J. de Jong^{2,18} & B. J. Trask¹²

* These authors contributed equally to this work.

1, Department of Pediatrics, University of Pennsylvania, The Children's Hospital of Philadelphia, 3516 Civic Center Boulevard, ARC 516, Philadelphia, Pennsylvania 19104, USA; 2, Roswell Park Cancer Institute, Elm and Carleton Street, Buffalo, New York 14263, USA; 3, National Center for Biotechnology Information, National Library of Medicine, Building 38A/Room 8N805, Bethesda, Maryland 20894, USA; 4, National Cancer Institute, NIH, Building 10/Room 12N214, Bethesda, Maryland 20889-5105, USA; 5, The Institute for

Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA; 6, Departments of Pediatrics and Human Genetics, Cedars-Sinai Medical Center, 8700 Beverly Boulevard, Los Angeles, California 90048, USA; 7, Computer Science Department, University of California Santa Cruz, 1156 High Street, Santa Cruz, California 95064-1077, USA; 8, Department of Biology, California Institute of Technology, Mail Code 147-75, Pasadena, California 91125, USA; 9, University of California San Francisco Cancer Center, Box 0808, San Francisco, California 94143-0808, USA; 10, Stanford University, Genome Lab, Mail Code 5120, Stanford, California 94305-5120, USA; 11, Sanger Center, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK; 12, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North C3-168, P.O. Box 19024, Seattle, Washington 98109-1024, USA; 13, Department of Molecular Biotechnology, University of Washington, Box 357730, Seattle, Washington 98195-7730, USA; 14, Departments of Obstetrics and Gynecology and Pathology, Brigham and Women's Hospital, Amory Lab Building 3rd floor, Boston, Massachusetts 02115, USA; 15, Department of Human Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA; 16, Joint Genome Institute-Los Alamos National Laboratory, MS M888 B-N1, P.O. Box 1663, Los Alamos, New Mexico 87545, USA; 17, Joint Genome Institute-Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 84-171, Berkeley, California 94720, USA; 18, Children's Hospital Oakland Research Institute, 747 52nd Street, Oakland, California 94609, USA; 19, Research Genetics, 2130 Memorial Parkway, Huntsville, Alabama 35801, USA; 20, Institute for Systems Biology, 4225 Roosevelt Way NE, Suite 200, Seattle, Washington 98105-6099, USA; 21, Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, 72 King William Road, North Adelaide, South Australia 5006, Australia; 22, Howard Hughes Medical Institute, Computer Science Department, University of California Santa Cruz, 1156 High Street, Santa Cruz, California 95064-1077, USA
† Present addresses: PanGenomics, 6401 Foothill Boulevard, Tujunga, California 91024, USA (U.-J.K.); Harvard Medical School, 240 Longwood Avenue, Cell Biology, Cambridge, Massachusetts 02115, USA (C.T.); Gene Logic, Inc., 708 Quince Orchard Road, Gaithersburg, Maryland 20878, USA (G.L.S.-O.).

Microarrays, DNA and RNA prep

Genomics — touched upon elsewhere in the issue — sneaks in here too.

Reagent kit and QC microarray slides

NEN Life Science Products www.nen.com

For Micromax or home-brewing applications

NEN's direct microarray reagent kit and QC slides are suitable for laboratories engaged in high-throughput differential gene expression research. The kit contains the key ingredients for direct cDNA labelling with fluorescence nucleotides. It provides sufficient reagents for five differential gene expression experiments. Non-mammalian, unlabelled control RNA and corresponding cDNA spotted on the QC microarray slides facilitate determination of sample labelling efficiency and provide internal controls for protocol development and validation. The products are suited to both home-made and Micromax microarray applications.

Reader Service No. 100



NEN's recently introduced reagent kit for direct cDNA labelling.

ArraySCOUT 2.0

Lion Bioscience www.lionbioscience.com

Enterprising software

The third release of the ArraySCOUT program for enterprise-wide gene expression data analysis has been upgraded with new clustering and visualization tools and an open application interface that allows users to add their own analysis methods. The program can handle any database system containing raw expression data and is fully integrated with two upcoming software modules for analysis of biological pathways and molecular interaction data.

Reader Service No. 101

Array Designer 1.10

Premier Biosoft www.PremierBiosoft.com

Rank and file

Array Designer is a program to process an unlimited number of sequences, using a powerful but fast algorithm to produce the best possible primers available within each target sequence, all in two to three seconds per sequence. Each primer and potential pairs are ranked for priming efficiency, with the best pairs/probes then supplied in a format ready for ordering from an oligo synthesis facility, importing in a central database such as Oracle, loading into a spotting robot, or printing, sorting and performing spreadsheet functions. Primers are screened for thermodynamic properties as well as secondary structures. A BLAST function launches cross-homology searches for each amplicon and probe to ensure specificity, with search results using the most current NCBI databases stored on the local hard drive. A demo version of the pro-

gram, which runs in Windows, can be found on the company's website.

Reader Service No. 102

Montage Genomics Kits

Millipore

www.millipore.com

Speedy sample prep

Two new kits — for 96-well plasmid miniprep and PCR clean-up — have been devised to speed up sample preparation. Simplified protocols eliminate lengthy bind and elute methods and centrifugation. The Montage plasmid miniprep kit recovers 96 clean DNA samples in half the time taken by traditional methods. One 25-minute protocol will yield DNA suitable for sequencing and transformation applications. The PCR clean-up kit avoids lengthy washing steps or organic solvents, minimizing the time needed to clean up PCR products using 96-well multiscreen plates.

Reader Service No. 103

Perfect gDNA Blood Mini Kit

Eppendorf

www.eppendorf.com

Bound to succeed

The 'Perfect gDNA' mini-kit is particularly suitable for isolation of high-molecular genomic DNA from difficult samples such as old blood, blood with a high erythrocyte content, or animal blood. The yields of genomic DNA obtained from a 1–200 µl blood sample are said to be close to the theoretical maximum, thanks to a combination of efficient lysis of base material and binding of the DNA to a specially developed membrane surface. After about 25 minutes, DNA is available directly following elution, with no further

precipitation steps necessary. Average fragment size is over 50 kb.

Reader Service No. 104

Phoretix Array²

Nonlinear Dynamics

www.nonlinear.com

Grid formation

This software enables the user to automatically detect the correct grid and subgrid structure in micro- and macroarrays by inputting the grid and subgrid proportions and shape of spot. In over 95% of cases, full analysis can be achieved by a single button press. Grid formats are automatically saved and there is an option for a rotatable area of interest to compensate for errors in image capture or to remove plate identifiers from the detection area. A manual method is available for specific applications.

Reader Service No. 105

RNAwiz

Ambion

www.ambion.com/rnaisol/

One-step RNA isolation

A single, homogeneous, organic reagent, RNAwiz is claimed to produce high-quality, intact RNA from tissues or cells in less than an hour. Using a single-step disruption/separation procedure, RNAwiz lyses cells and dissolves cellular components while protecting RNA integrity. Any amount of starting material can be accommodated and the system effectively disrupts tissues rich in polysaccharides, fatty acids or proteins. Total RNA isolated is free of protein and most DNA contamination and is suitable for all downstream applications.

Reader Service No. 106



MyArray: DNA ready for custom microarrays.

MyArray

Invitrogen www.invitrogen.com

Knock the spots off with this DNA

This PCR product can be amplified from any number of cDNA clones or libraries and delivered in 96-well microtitre plates for researchers wishing to produce their own microarrays. The cDNA clones are arrayed, amplified, purified and shipped ready to be spotted onto glass slides. Absorbance readings for glycerol stock cultures, gel images of the amplified PCR product and an annotated data file are available to users at the company's secure website.

Reader Service No. 107

Wizard

Promega www.promega.com

Magnetic DNA purification for the gourmand

Billed as the first commercially available system specifically designed for purification of DNA from food, Wizard uses Magnesil paramagnetic particle technology to purify nucleic acids by magnetic separation. PCR-quality DNA suitable for detection of GMO in food can be purified simply and quickly from a cleared lysate. The system can be used to purify genomic DNA from a variety of foods, including corn and soya seeds, processed foods such as cornmeal, cornstarch, soya flour, cornflakes, soya milk and food samples with low DNA content or technical obstacles, such as soya lecithin and vegetable oils. The technology can be automated for large-scale processing.

Reader Service No. 108

Taq DNA polymerase

Pierce www.piercenet.com

Hot new enzyme

The Pierce Taq DNA polymerase is a recombinant form of the native enzyme isolated from *Thermus aquaticus* and expressed in *Escherichia coli*. It is licensed for use in PCR and is suitable for amplification of DNA fragments from a wide variety of templates, including

plasmid DNA, genomic DNA and cDNA. The high purity and thermostability of the enzyme allow single-copy genes to be efficiently amplified. A product instruction book can be downloaded from the company's website.

Reader Service No. 109

Nucleon

Tepnel www.tepnel.co.uk

Genomic DNA purification in 30 minutes

Nucleon phenol-free, ready-to-use blood kits facilitate rapid extraction and purification of high molecular weight genomic DNA from blood and cell cultures. They can purify DNA in 30 minutes and, since the extraction system is not based on column format, they are readily scaled up to accommodate large sample volumes. The recovered DNA is suitable for PCR, restriction digest, blotting and hybridization.

Reader Service No. 110

CDNA Integrity Kit

KPL www.kpl.com

Full-length demonstration

KPL's new cDNA integrity kit provides a method of evaluating cDNA preparations for the presence of full-length and extended transcripts prior to use. Results can be obtained from any single- or double-stranded cDNA from human, mouse or rat cells and tissue. The kit contains six primer sets that amplify specific regions of four commonly expressed genes (clathrin, GAPDH, S6 and L3). Universal PCR conditions provide amplification of all six primer sets in a single cyclo run. Applications include library construction, RT-PCR, microarrays, RAGE, and differential display.

Reader Service No. 111

TAP Express

Gene Therapy Systems, Inc.

www.genetherapysystems.com

Active PCR fragments, quickly

The new Transcriptionally Active PCR (TAP) Express Rapid Gene Expression kit from GTS is claimed to be the first commercially available system for rapid construction of transcriptionally active PCR fragments for expression in mammalian cells. This expression system allows for time and cost-savings for genetics studies by eliminating traditional cloning, transformation and plasmid preparation procedures. Only two PCR steps are involved with TAP Express, and the gene-of-interest is ready for expression in 1 day.

Reader Service No. 112

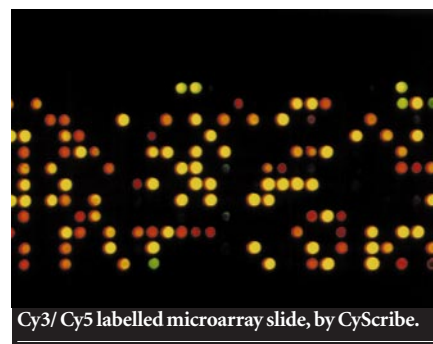
CyScribe

Amersham Pharmacia Biotech

www.apbiotech.com/cydye

First-strand cDNA labelling kit

Designed for the preparation of highly fluorescent first-strand cDNAs for use as probes



Cy3/Cy5 labelled microarray slide, by CyScribe.

in microarray hybridizations, this kit includes optimized reagents and protocols to generate cDNA probes labelled with Cy3 or Cy5 dCTP or dUTP deoxynucleotides. The manufacturers claim that CyScript, their modified reverse transcriptase, achieves more efficient and even dye incorporation than other reverse transcriptase enzymes, resulting in brighter probes and more sensitive results. A microarray hybridization buffer is included to ensure low backgrounds.

Reader Service No. 113

These notes are compiled in the Nature office from information provided by the manufacturers.

Where the jobs are

Genomics and bioinformatics companies in North America
p962

Future paths

Current roles suggest what kind of opportunities lie ahead
p963

Additional skills

An ability to write software may help biologists find creative work
p963

Silicon heaven

Computer simulation could be a model career for biologists
p964

Labs and companies seek their niches as work continues after the draft

Specialization may be the key to success as divisions between big and small genomics centres continue to grow, say Potter Wickware and Paul Smaglik.

Now that the human genome has been sequenced in draft form, will there be less work in sequencing centres? Far from it, says Francis Collins, director of the US National Human Genome Research Institute (NHGRI). The biggest sequencing labs will switch from human to a list of other organisms. Smaller labs will focus on finishing and annotation. The output of both will crank the demand for bioinformaticians even higher. The common denominator? A strong computational background, lab heads say.

Opportunities, big and small

The big-science model, in which large sequencing centres dominate production of sequence data, will continue at least in the near term, even though large-scale sequencing of the human genome is now drawing to a close. Those centres — at Washington University in St Louis; the Sanger Centre in Hinxton, Cambridgeshire; the Whitehead Institute at the Massachusetts Institute of Technology; Stanford University; and the US Department of Energy's Joint Genome Institute (JGI) in Walnut Creek, California (a consortium of university and national lab sequencing centres) — will be turning their high-throughput capacities towards other organisms.



Get your motor running: there's plenty of work for self-starters, says Francis Collins.

A vast supply of other genomes beckons, representing a mostly untapped resource with huge potential payoffs for agriculture, environmental remediation and medicine.

Dan Rokhsar, director of bioinformatics at the JGI, says that advances in sequencing and assembly mean 95% of a typical microbial genome can now be sequenced in a day and a half. The JGI sequenced 15 last October, and proposes to do three times that number each year, in a set of 'microbial marathons'.

But the big centres will not eclipse the smaller labs — as long as they do not try to compete in the high-throughput sequencing game. "The smaller centres will need to develop themes, as it is unlikely that they can match the largest centres on a pure cost-per-base-pair basis," Collins says.

'Finishing', or filling in the many gaps in the draft version of the human genome, is a growth industry, he says. Only about a quarter of the genome can be considered finished, and the publicly funded project is committed to completing the rest by the end of 2003. Maynard Olson's University of Washington lab is actively engaged in closing gaps in the human genome that result from



Following the growth of data

In the early 1980s David Haussler worked with pattern recognition on

some of the earliest available DNA, the phage ϕ X174.

"A single seamless career working with viruses and ribosomal binding sites would have been possible, but the lack of data was frustrating," he says. It took an inordinate amount of bench work to get more, so he felt

it was more logical to go off and develop fundamentals in statistics and machine analysis methods.

"If the data had been there I would have stayed with bioinformatics, because the early work we were doing with promoters was so exciting," recalls Haussler, who is now director of the Center for Biomolecular Science and Engineering at the University of California at Santa Cruz.

Later, of course, the problem

became, if anything, too many rather than too few data. When he returned to sequence analysis, Hidden Markov Modelling — a robust set of pattern recognition methods that was developed by his group — was in place to help deal with the task.

"We're just beginning to get to know the genes. There's a huge amount to be discovered as we push on to link genes with disease and basic molecular biology," Haussler says. **P. W.**

careers and recruitment

the high-throughput approach. But the sheer size and difficulty of the task means ample opportunities for latecomers.

Other labs may specialize further, finishing parts that resist being sequenced by machines. Regions near telomeres and centromeres may be well suited to small labs, as these stretches of DNA require time-consuming manual techniques to decipher, says David Haussler, director of the Center for Biomolecular Science and Engineering at the University of California at Santa Cruz.

Funding abundance

Specialized labs not directly associated with sequencing will also emerge. The NHGRI received a 15% increase for the fiscal year 2001 (which started last October), increasing its budget to \$382 million. Most of the increase will go towards the new Centers of Excellence in Genomic Science programme.

This will fund multidisciplinary centres focusing on genome-wide analysis and technology development. Although applications will not be reviewed until May, the programme is likely to create several multi-million-dollar centres for analysing gene function and gene expression; studying population genetics and sequence variation; and performing comparative genomics and complex trait analysis, among other things. Most will require scientists with sophisticated computational skills.

Other agencies also benefit from federal largesse. The Department of Energy allocated \$117 million to genomics in 2001 — up 14% from last year — and the National Science Foundation, which got a rise of \$500 million this year to \$4.4 billion, is supporting genomics-related research. Funding for plant research, still far behind human-related work, is also on the rise, with a current US Department of Agriculture genomics budget of \$85 million, up from \$79 million last year.

The states are being generous, too. California's new Institute for Bioengineering, Biotechnology and Quantitative Biomedical Research, announced in December, will receive \$100 million in public and \$200 million in private donations. It will be based at the University of California at San Francisco's new Mission Bay campus, with major components at Santa Cruz and Berkeley, and will tie into Berkeley's \$500 million Health Sciences Initiative, which seeks to advance health science research through multidisciplinary collaborations.

Bioinformatics needs

All these initiatives will need people with the computational skills to analyse increasingly complex data sets. But bioinformaticians, like smaller sequencing centres, may find that it pays to specialize.

Mark Gerstein, who leads a bioinformatics group at Yale University, sees opportunities for the independent researcher to carve a niche in a particular type of promoter or

Genomics and bioinformatics companies in North America

Company	Activity	URL
Affymetrix*	Expression chips, analysis services;	www.affymetrix.com
AP Biotech*	Comprehensive drug development	www.apbiotech.com
Base4	Pharmatrix database; pharma and IT consulting	www.basefour.com
Caprion Pharmaceuticals	High-throughput cell maps, proteomics	www.caprion.com
Celera*	Databases, genotyping, target discovery	www.celera.com
Cellomics	Whole-cell assay chip	www.cellomics.com
CuraGen*	SNPs, expression, low abundance genes, drug-induced changes in gene expression	www.curagen.com
Deltagen	Functional genomics	www.deltagen.com
Digital Gene Technologies	TOGA system correlates expression with anatomy; also has LIMS	www.dgt.com
DNA Sciences (was Kiva Genetics)	Gene Trust patient database	www.dna.com
DoubleTwist	Web-based software "agents"	www.doubletwist.com
Exelixis	Pathway and target finder software	www.exelixis.com
Gemini Genomics*	Genotyping, SNPs	www.gemini-genomics.com
First Genetic Trust	Database; patient information encryption	www.firstgenetic.net
Genaissance	Population genomics, "personalized medicine"	www.genaissance.com
Gene Logic*	Expression analysis	www.genelogic.com
Genicon	RLS method of labelling & detecting biomaterials	www.geniconsiences.com
Genomica*	Software for family studies, epidemiology	www.genomica.com
Genomics Collaborative	SNP genotyping	www.getdna.com
Genomics Institute (a division of Novartis)	Comprehensive genomics, proteomics; mouse genetics	www.gnf.org
Genomic Solutions	Biochips & arrays	www.genomicsolutions.com
Human Genome Sciences*	Sequencing, expression analysis, proteomics, preclinical & clinical testing	www.hgsi.com
Hyseq	Genomics, high-throughput sequencing	www.hyseq.com
IBM*	"Blue Gene" supercomputer, computational biology group	www.research.ibm.com/topics/serious/bio/
Integrated Genomics	Microbial genomes, pathways	www.integratedgenomics.com
Incyte*	Databases, software	www.incyte.com
Informax*	Vector NTI software	www.informax.com
LabBook	Genomic XML browser	www.labbook.com
Lexicon Genetics	Biochips & arrays; OmniBank knockout mouse clones	www.lexgen.com
Lynx Therapeutics	SNP genotyping; MegaClone expression method	www.lynxgen.com
Maxygen	"Gene breeding" directed evolution compounds	www.maxygen.com
Motorola BioChip	Expression arrays	www.motorola.com/biochipsystems/
Molecular Simulations	Software	www.msi.com
Myriad Genetics*	Sequencing, proteomics	www.myriad.com
Nanogen*	Biochips & arrays	www.nanogen.com
NetGenics*	Software	www.netgenics.com
Orchid Bioscience	SNP genotyping	www.orchid.com
Paradigm Genetics	SNP genotyping	www.paragen.com
Phylos	"HIP" protein chip	www.phylos.com
Promega	SNP genotyping	www.promega.com
Prospect Genomics	Structure prediction	www.prospectgenomics.com
Proteome (division of Incyte)	Worm, yeast databases	www.proteome.com
Qiagen Genomics	SNP scoring	www.qiagen.com
Rosetta Inpharmatics*	Expression analysis software	www.rosetta.com
Senomyx	Smell & taste genes	www.senomyx.com
Sequenom*	SNP analysis	www.sequenom.com
Silicon Genetics	Array analysis software	www.sigetics.com
Spotfire	Analysis "siftware"	www.spotfire.com
Syrrx	High-throughput protein structure prediction	www.syrrx.com
Structural Bioinformatics	Patient-specific structural variants	www.strubix.com
Structural Genomix	Proteomics, structure	www.stromix.com
Zyomyx	Protein biochips	www.zyomyx.com

(* = public company)

structural motif, then carry out the analysis on the entire genome.

"It doesn't require a huge amount of apparatus, and the full annotation of the genome is such a huge task that there's room for everyone," he says.

How many people practise the hard-to-define trade is hard to pinpoint, but a good estimate is attendance at the annual Intelligent Systems for Molecular Biology conference, which began with just 200 in 1993. The figures started zooming up three years ago,

rising to 1,300 last year. Worldwide, the number is probably two or three times greater. Bioinformatics may be a small field, but few areas can boast such a steep rate of gain.

The jobs are spread across industry, academia and government labs. Enterprises ranging in size from Fortune 100 corporations to tiny boutique companies all have the 'Help Wanted' sign up. At least 50 companies devoted to genomics, proteomics or bioinformatics are doing business in North America (see table). Many corporations have their own groups and list openings on their web pages; collective listings are also found on bulletin boards such as <http://scijobs.org> and <http://www.genomeweb.com>.

Major pharmaceutical companies are a rich source of bioinformatics jobs. Terry Gaasterland, director of the Laboratory of Computational Genomics at Rockefeller University in New York, thinks they lead in

bioinformatics analysis. They have far more data than anyone else, and because of the many specific tasks they address they are ahead with their methods as well. "For example, pharmas have much better methods of dealing with Affymetrix expression data than the academic sector," Gaasterland says.

Reaching this stage of the human genome project has created ample opportunities for people adept at adapting — and adapting to — new technologies and computational approaches, says Collins. "If you're bright, motivated, creative, and can set up automated PCR and write a Perl script, there's a promising future for you in genomics." ■

Potter Wickware is a science writer in San Francisco,

Paul Smaglik is editor of *Naturejobs*.

SNP consortium ♦ <http://snp.cshl.org>

Centers of Excellence in Genomic Science

♦ http://www.nih.gov/Grant_info/Funding/

Research/CEGS_synopsis.html

piece jigsaw with no guiding picture, using pieces of jigsaw characterised by only four shapes. Even the private sequencing venture run by Celera Genomics has turned to the publicly published maps as a guide to assembling their sequence.

Mapping jobs at the Sanger Centre, in Hinxton, Cambridgeshire, are less in demand now that the genome has reached draft quality. At the height of the human mapping efforts, there were some 60 to 70 dedicated mappers. Now that the Sanger has turned to sequencing pathogens and smaller model organisms such as zebrafish, the centre only needs 18 to 20 mappers. They are focusing on verifying the draft and finished version of the human genome, and on learning as much as possible from its raw data. Panos Deloukas, a project leader and one of the key players in producing the physical map of the human genome, is studying the occurrence of disease in three different populations (African Americans, Asians and Caucasians) and matching disease phenotypes to genetic mutations.

The kind of work Deloukas does requires a PhD and experience as a postdoc. Universities, institutes and industry are finding that candidates with these qualifications are in short supply. Says Michael Ashburner, joint head of the European Bioinformatics Institute (EBI): "There are very few good people at a senior level, and there are very good posts we cannot fill." Yet it is not just those with PhDs who find work at sequencing centres.

Clone rangers

Preparation of the draft published today was generated using cloned contigs. In this approach, the DNA is fragmented into sequences that contain markers from the published genetic and physical maps, and inserted into bacteria. Each clone is then randomly broken into smaller segments which are sent to the shotgun teams for sequencing, then to finishing and annotation. Finally, the finished contigs are placed on the master map in the position defined by their markers.

The range of qualifications needed for shotgun sequencing, finishing and analysis vary considerably. Bev Mortimore, a shotgun team leader, was one of the original 17 staff to accompany John Sulston's team when it moved to the Genome Campus at Hinxton (the Sanger now employs 570 people). She says that the sequencing work can be repetitive and she prefers not to hire graduates who find the routine tedious. It is suited to intellectually curious school-leavers (perhaps with GCSEs, perhaps 'A' levels) who learn on the job.

Finishing touch

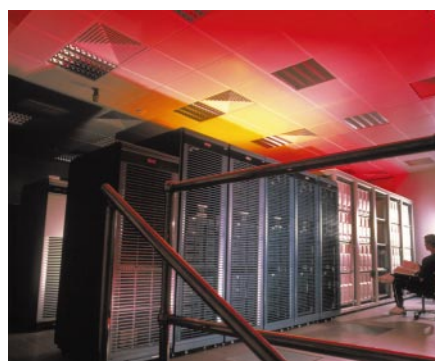
Once each fragment is sequenced, it is reassembled by running computer programs that seek out areas of overlap. But there are always gaps where it has not been possible to identify the sequence correctly and find areas

Current role suggests the shape of future work opportunities

Mappers, cloners, sequencers, finishers and annotators — each of the five major sequencing centres and the nine responsible for smaller portions of the genome employ such staff. How do their tasks fit together? And, now that the project is racing to the finishing stages, what will people at centres worldwide turn their attention to? One way to get an idea is to look at their role in the human genome project up to now.

Before they could proceed into the uncharted territory of human high-throughput sequencing, geneticists needed maps. First, scientists created a genetic map, which plotted the approximate location on each chromosome of genetic features such as a gene or particular base sequence, determined by studying genetic material from small populations of families. Then they sketched a physical map, using molecular biology to determine the location of specific sequences.

Combining, then refining, those maps



Computers will play an increasing role in biology.

has been critical to the production of the draft human genome published today. By locating a number of DNA sequences on the chromosomes, biologists provided themselves with a framework on which they could place sequence data. Without maps, accurately positioning DNA sequences would have been akin to assembling a 3-billion-

Who makes the best bioinformaticians?

Bioinformatics careers can be divided into two paths: developing software, and using it. The field, catalysed by the rapid accumulation of genomic data, has attracted attention as a salvation for jobs in biology. But that sentiment may not provide an accurate assessment of job opportunities, at least for career prospects on both paths. For example, InforMax, one of the largest bioinformatics companies in the United States, generally doesn't hire biologists-turned-programmers, says Alex Titomirov, chairman and chief executive officer of the company, based in North Bethesda, Maryland.

InforMax has about 95 programmers, almost all of whom come from a maths, physics or computer-science background. Titomirov says it is "much easier" to teach people with those skills about biology than to teach biologists how to code well. However, as the company turns to developing software to handle functional genomics and protein data, it may draw on more biologists to help design new software modules. **P.S.**

of overlap. This is where the finishers, who are mostly graduates with some element of biology in their degrees, take up the job. "No matter how hard you try," says Karen Barlow, one of five finishing team leaders, "there are places where you can't figure out what's going on. Maybe a length of sequence physically doubles back on itself, and you need to go in and break it into smaller pieces."

Each finisher juggles between 10 and 25 gaps that need filling. They decide what experiments – wet biology, not *in silico* – are needed to understand why there is not a good match between two pieces of DNA. Barlow has worked her way up from finisher via senior finisher to one of five team leaders during her six years at the Sanger.

Annotation jamboree

After finishing (99.99% confidence that the bases are correctly labelled and positioned), the sequence goes to an annotator. Their task is to provide as much scientific description of the sequence and its function as possible. It is during the annotation and analysis that the biological purpose of the mapping and sequencing effort starts to become clear. All the information is placed in the DNA databank at the European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory, and mirrored in other public databases in Japan and the United States.

Within the Sanger, a number of projects provide information to the annotators. One example is Pfam, which stands for 'protein family'. Alex Bateman, one of the Pfam group, says: "As the sequence data come in, we look to see what it is coding for and classify the protein according to family (say the trypsinogen family). For this job you need to be skilled with bioinformatics tools and have good biological knowledge so that you can evaluate the plausibility of the answer that pops out of the computer."

Applying the mathematical sciences to biology is "a new, young field and no-one knows where it's going," says Stephen Bentley,

Tips for sequence centre job-seekers

If you want to get a job at a sequencing centre, first check out its website in detail. "It amazes me how often graduates say how much they want to work at the Sanger, but they haven't bothered to look at our website," says Bev Mortimore, shotgun team leader. "It's my trick question, to ask them what they think of the site."

Get an MSc in bioinformatics, says Alex Bateman, who works on classifying proteins into families. You're likely to be poached straight from the course.

An MSc in bioinformatics alone, though, isn't enough if you want to develop new computational tools, says Tim Hubbard, head of human sequence analysis (and a physicist by training) at the Sanger Centre. You need to demonstrate that you've downloaded and installed Linux (a new Unix-style operating system), say, and know your way around an operating system. **H.G.**

Five major sequencing centres are responsible for producing the draft human sequence. They have links that take you to all relevant sites.

The Sanger Centre
♦ www.sanger.ac.uk

Washington University Genome Sequencing Centre in St Louis
♦ <http://genome.wastl.edu/gsc>

Whitehead Institute in Cambridge, Massachusetts

♦ www-genome.wi.mit.edu

Baylor College of Medicine
♦ www.hgsc.bcm.tmc.edu

The Joint Genome Institute of the US Department of Energy
♦ www.jgi.doe.gov

an annotator in the pathogen-sequencing unit at the Sanger. Bentley has a BSc in applied biology, worked on *E. coli* for his PhD and joined the Sanger from a postdoc position at the University of Cambridge.

One of the fun aspects of the work, he says, is seeing the new things that roll past. For example, he remembers spotting one gene that he knew had never been seen in a prokaryote before. He looked for sequences coding for a similar gene in eukaryotes and found that the gene was implicated in disease. He then called the Canadian scientists involved to tell them about the new gene in prokaryotes. "What I love is that you don't have to be protective about what you find, you can call a scientist and alert them to something that might be significant."

Functional future

In the coming years more genomics groups exploiting the expertise of sequencing centres will want to be located near one of the 16 sequencing centres worldwide. The proximity will be helpful when these functional genomics

centres need to resequence genes in order to verify their biological function.

Already the UK's Human Genome Cancer Project has moved into the Sanger, where the researchers sit cheek by jowl with sequencers, annotators and computational biologists. The project's aim is to study the contribution to tumour formation of genetic variations. Simply being in the same building makes their work very much easier, says Sarah Edkins, a senior research assistant. In the past six months they have developed 1,000 cancer lines.

So, with additional organisms to sequence and the resequencing needed to verify the computer-based assignment of function mappers, there will still be work for shotgun sequencers, finishers, annotators and analysts. However, the numbers required to fill positions in each discipline is changing to meet the shift in sequencing focus, increased automation at every step and the development of new computational tools. ■

Helen Gavaghan is a freelance science and technology journalist based in Hebden Bridge, West Yorkshire.

Biology moves into the silicon stage

First there was *in vivo* biology, then *in vitro* and now the discipline is moving *in silico*. Or, to paraphrase US political strategist James Carville, whose campaign slogan helped put former US President Bill Clinton in power, the watchwords for future success in biology could easily be: "It's the computing, stupid". Biologists will have to be adept at least in handling the tools of bioinformatics, and, for true insight, they will be better placed learning the informatics needed to create at least some tools for themselves. Says Tim

Hubbard, head of human sequence analysis at the Sanger Centre in Hinxton: "If you don't, you are at the mercy of the developers."

One man who has seen the transition to the *in-silico* world is Graham Cameron, a loquacious and amiable Scot, who was the sole person looking after the newish database at the European Molecular Biology Laboratory (EMBL) in 1982. He was one of the prime movers in the formation of the EMBL's European Bioinformatics Institute (EBI), which he

now heads along with geneticist Michael Ashburner.

The EBI has five major databases that should aid in understanding the raw data of the human genome. These are: the EMBL-DNA sequence database; protein sequences (jointly with the Swiss Bioinformatics Institute), ENSEMBL (human and mouse data — a joint effort with the Sanger for automatic characterization of gene function); protein structure; and a gene-expression database.

Cameron is head of services, but in

recent years his main role seems to have been fighting for funds to ensure that European science and business do not lag behind the United States in the genomics revolution. Almost every branch of the Institute needs to grow, he says, if it is to serve the scientific community well in the next critical few years when the world's scientists — academic and industrial — will throw themselves at these data in search of major scientific breakthroughs and commercial success. **H.G.**

♦ <http://ebi.ac.uk>